Glottometrics 17
2008

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

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

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

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 17 (2008), Lüdenscheid: RAM-Verlag, 2008. 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 17 (2008) **ISSN 2625-8226**

Contents

History of Quantitative Linguistics

Glottometrics 17, 2008, 1-7

Verteilung von Wortlängen in englischen Spam-E-Mails

Thomas Jahn, Annika Uckel[1]
Universität Göttingen

Abstract. The present paper deals with word length in twenty English spam e-mails and is part of the Göttingen *Projekt Quantitative Linguistik*. In order to get a homogenous text sample, all of the e-mails are impersonal mails with a length between 75 and 304 words. We show that the *Hirata-Poisson* distribution appears to be a good model in the majority of texts analyzed. This paper tries to make a contribution to the hypothesis that word length in texts abides by certain distribution laws.

Keywords: Word length, English, spam

1. Vorbemerkungen

In der vorliegenden Arbeit wird die Wortlängenverteilung in englischsprachigen Spam-E-Mails untersucht. Als Datengrundlage dienen 20 Texte, die im Zeitraum von April bis August 2007 gesammelt wurden. Die Untersuchung soll zur Erweiterung der empirischen Basis der in den letzten Jahren vermehrt durchgeführten Untersuchungen zur Verteilung von Wortlängen in Texten dienen. Anhand der vorliegenden Arbeit soll die Hypothese überprüft werden, dass die Verteilung von Wortlängen in Texten einem gesetzmäßigen Verlauf folgt und mit einer von wenigen Verteilungsfunktionen modellierbar ist (Wimmer et al. 1994; Wimmer & Altmann 1996), wie im Göttinger Projekt zur Quantitativen Linguistik (Best 1998) bereits festgestellt wurde.

2. Textauswahl

Aufgrund der Erfahrungen mit Wortlängen im Englischen, gemessen nach der Zahl der Silben, werden bei der Selektion und Bearbeitung der Texte die Erkenntnisse der bereits vorhandenen Wortlängenanalysen berücksichtigt (vgl. Best 2001, 2).

Bei Spam-E-Mails handelt es sich aufgrund ihrer Kürze, der Bearbeitung durch viele verschiedene Autoren und durch die zahlreichen unterschiedlichen Arten und Funktionswiesen von Spams um eine sehr heterogene Textsorte. Um für diese Untersuchung ein möglichst homogenes Textkorpus zu erhalten, wurden die Spam-E-Mails im Vorfeld anhand nachfolgender Kriterien selektiert: das Textkorpus besteht nur aus so genannten „unpersönlichen Spam E-mails“ (vgl. Schmückle, Chi 2004: 40), welche im Gegensatz zu herkömmlichen Privat-E-Mails oder „persönlichen Spam-E-Mails“ auf Anredeformeln verzichten und keine Andeutung einer persönlichen Beziehung vom Sender zum Empfänger machen. Texte mit hohem Fremdwortanteil sowie vielen typologischen Besonderheiten wurden bewusst vermieden, da diese Sonderfälle schnell zu Modellierungsproblemen führen können. Da es sich hier vermutlich um die erste derartige Untersuchung von Spams handelt, sollten solche Komplikationen zunächst vermieden werden. Die durchschnittliche Textlänge liegt bei 126 Wörtern, wobei der längste Text 304 und der kürzeste 75 Wörter umfasst.

[1] Address correspondence to: thomasjahn@gmx.com or Annika.Uckel@gmx.de

3. Definitionen

Die Wortlängen in den Spam-E-Mails wurden in Silben gemessen. Bei der Bestimmung der Wortlänge anhand von Silben ist es erforderlich, die Begriffe „Wort“ und „Silbe“ zu definieren.

3.1. Wort

Das Wort, welches den Untersuchungsgegenstand dieser Arbeit darstellt, wird als kleinste orthographische Einheit verstanden. Dem Begriff des Wortes liegt die Definition von Best zugrunde, der das Wort als „ununterbrochene Graphemkette“ kennzeichnet, „die von Leerstellen oder Interpunktionszeichen begrenzt wird“ (Best 2006, 24). In der vorliegenden Arbeit werden Wörter, die durch Bindestriche, Apostrophe oder Querstrich verbunden sind, als jeweils ein Wort gezählt.

3.2. Silbe

Die Silbe wird bei Best als „Sprecheinheit, die bei langsamem Sprechen hervortritt“ definiert (Best 2005: 5). Die Silbe besteht aus einem Silbenanlaut, einem Silbenkern und einem Silbenauslaut. Obligatorischer Bestandteil ist nur der Silbenkern, bestehend aus einem Vokal oder Diphthong (vgl. Best 2005: 5). Jede Silbe beinhaltet demzufolge nur einen Vokal oder Diphthong, der sich im Zentrum der Silbe befindet. Neben den Diphthongen existieren in der englischen Sprache mehrere Wörter, die Triphthonge enthalten (z.B. fire). Statt diesen Dreilaut als zwei Silben zu zählen, also Diphthong plus Vokal, werden Triphthonge in der vorliegenden Arbeit als Kriterium für eine Silbe gewertet. Die Silbenanzahl eines Wortes hängt somit von der Anzahl an Vokalen, Diphthongen oder Triphthongen ab.

Die phonologische oder phonetische Silbe darf nicht mit der orthographischen Einheit verwechselt werden, die aus der Silbentrennung hervorgeht (vgl. Best 2005: 5).

4. Bearbeitung der Texte

Als zum Text zugehörig wird nur der fortlaufende Text ohne Adress- und Betreffzeile, Hyperlinks oder Fußnoten gezählt. In Klammern gesetzter Text wird berücksichtigt.

Insgesamt wird der Text gemäß seiner phonologischen Realisation ausgewertet; das bedeutet, das Wort wird als phonologisch transkribiert gedacht. Bei der Bestimmung der Silbengrenzen wird auf eine phonologische Umschrift der Texte verzichtet. In Zweifelsfällen wurde das *Cambridge Pronouncing Dictonary* (2006) zur Bestimmung der Silbenanzahl herangezogen.

Gebräuchliche Abkürzungen werden ihrer ausgesprochenen Form nach gezählt, so beispielsweise „Mr.“ als „Mister“. Bei ungebräuchlichen Abkürzungen, wie bei „HGH“, wird jedes Graphem als einzelne Silbe ausgezählt. Ebenso wird mit Symbolen, wie beispielsweise dem Dollar-Zeichen, verfahren, die gemäß ihrer phonologischen Realisation gewertet werden. Bei Elisionen oder anderen Aussprachevarianten, die bei der britischen Aussprache unter Umständen auftreten, wird jeweils die längere Variante gewählt. Zusammengezogene Formen, welche in der englischen Sprache häufig Verwendung finden, werden in ihrer Langform ausgewertet, so beispielsweise „haven’t“ als „have not“. Die Markierung des Genitivs im Englischen durch „’s“ wie etwa bei „the body’s organs“ wird als zum Wort zugehörig betrachtet.

Zahlen werden nicht als orthographische Einheit im Sinne der Definition aufgefasst. Bei Zahlen wird wie bei Ammermann (2001, 65) eine Einteilung in Tausender-, Hunderter, Zehner- und Einerschritte vorgenommen. Durch die Unterteilung der Zahlen wird verhindert, dass unnatürlich hohe Silben- beziehungsweise Wortlängen entstehen, die das Ergebnis der Auswertung negativ beeinflussen könnten. Bei Zahlen unter eins wird der Punkt gemäß seiner phonologischen Realisation betrachtet, so beispielsweise „0.56“ als „zero“ „point“ „five“ „six“ ausgewertet. Bei Datumsangaben wird jede Zahl als einzelnes Wort angesehen.

Um die Auszählung zu erleichtern, wurden den verschiedenen Silbenanzahlen Farben zugeordnet, mit denen die entsprechenden Wörter im Text markiert werden. Für jede Mail wird in einer Strichliste die Häufigkeit von *x*-silbigen Wörtern (*Nx*) festgehalten. Die Auswertung der Strichlisten stellt die empirische Datengrundlage der Untersuchung dar.

5. Modellierung der Daten

Um herauszufinden, welches der in Frage kommenden mathematischen Modelle am besten an die Daten angepasst werden kann, wurden die Ergebnisse der Auszählung mithilfe des Altmann-Fitters (1997) getestet. Dabei erwies sich die Hirata-Poisson-Verteilung als am besten geeignet, um an die Daten angepasst zu werden. Die Formel der 1-verschobenen Hirata-Poisson-Verteilung lautet wie folgt:

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

wo [z] die größte ganze Zahl ≤ z ist (vgl. Stark 2001, 155). Diese Verteilung wurde des Öfteren in west-europäischen Sprachen gefunden. Sie eignet sich besonders deswegen für diese Zwecke, weil sie aus einer Kombination von zwei Verteilungen besteht, nämlich entweder aus der *Zusammensetzung* der Binomialverteilung und der Poissonverteilung bzw. der Poissonverteilung und der Normalverteilung oder aus der (Fellerschen) *Verallgemeinerung* der Poissonverteilung durch die 1-verschobene Null-Eins-Verteilung oder schließlich aus der *Faltung* der Poissonverteilung mit der Dublet-Poissonverteilung. Diese Provenienz zeigt, dass man im Grunde von der Poissonverteilung ausgeht und sie durch andere Verteilungen modifiziert, die durch bestimmte Randbedingungen hervorgerufen werden. Die genaue Form dieser Randbedingungen wurde noch nicht erforscht.

Die Verteilung muss in 1-verschobener Form verwendet werden, da es keine nullsilbigen Wörter gibt.

6. Anpassung des Modells

Die folgenden Tabellen enthalten die Anpassungen der 1-verschobenen Hirata-Poisson-Verteilung an die 20 untersuchten Spam-E-Mails. Als zufriedenstellend gilt die Anpassung, wenn $P \geq 0.05$ oder $C \leq 0.01$. Toleriert werden noch Anpassungen mit $P \geq 0.01$ und $C \leq 0.02$. Der Diskrepanzkoeffizient C wird bei derart kurzen Texten nur dann benötigt, wenn durch Zusammenfassung von Längenklassen null Freiheitsgrade gegeben sind. Das ist nur bei den Texten 8 und 17 der Fall.

Die Zusammenfassung von Längenklassen wird in den Tabellen mit einem senkrechten Strich (|) markiert. In den Tabellen bedeuteten:

x = Länge
n_x = beobachtete Häufigkeit
NP_x = berechnete Häufigkeit
a, b = Parameter Hirata Verteilung
X^2 = Chiquadrat
P = Überschreitungswahrscheinlichkeit des Chiquadrats
FG = Freiheitsgrade

	Text 1		**Text 2**		**Text 3**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	55	54.1310	64	64.1486	86	82.8265
2	23	22.8989	28	25.6467	42	42.0971
3	15	17.2912	5	7.9600	18	24.1650
4	7	5.9487	3	2.2447	7	8.6571|
5	4	3.7302			8	3.0645|
6					1	1.1898|
∑	104		100		162	
	a = 0.6530, b = 0.3522 X^2 = 0.52, FG = 2 P = 0.77		a = 0.4440, b = 0.0995 X^2 = 1.57, FG = 1 P = 0.21		a = 0.6708, b = 0.2424 X^2 = 2.43, FG = 1 P = 0.12	

Text 1: Working communities are diverse, reflecting the nation's racial and ethnic demographics. (Tim, <vvs@canada.com>)
Text 2: Amazing low rates! („Cherry Alfaro", <Maileeuwenhoekdraftsmen@spondylitis.org>)
Text 3: Serious job offer for serious people. (Mark A Steliga, <ffnha@gilbarco.com>)

	Text 4		**Text 5**		**Text 6**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	47	46.1161	50	48.4156	54	54.2284
2	18	17.7597	18	17.9093	18	17.9283
3	6	8.0874	10	13.2191	8	6.7942
4	3	2.2366|	5	4.0730	1	2.0492
5	1	0.8002|	3	2.3830		
∑	75		86		81	
	a = 0.4863, b = 0.2081 X^2 = 0.86, FG = 1 P = 0.35		a = 0.5745, b = 0.3562 X^2 = 1.21, FG = 2 P = 0.55		a = 0.4012, b = 0.1760 X^2 = 0.75 FG =1 P = 0.39	

Text 4: over 1,500,000 bottles sold worldwide (<„Ruth Lucas", goldston4@fone.net>)
Text 5: A little secret to make you private life more interesting! (<Williams, fgzhe@industrial.com>)
Text 6: Jeremiah – Viagra for you! (<„Jeremiah Parrish", dncxy@boneybooks.com>)

	Text 7		**Text 8**		**Text 9**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	123	123.000	62	62.0000	94	94.0044
2	29	29.0000	18	18.0000	44	43.9983
3	9	9.0513	24	17.8638|	14	13.9129
4	2	1.9487	2	8.1362|	4	4.0844
∑	163		106		156	
	a = 0.2816, b = 0.1626 X^2 = 0.0016, FG = 1 P = 0.97		a = 0.5363, b = 0.4587 C = 0.0000		a = 0.5065, b = 0.0759 X^2 = 0.0023, FG = 1 P = 0.96	

Text 7: Some momentum news? (<„Juli Kooh“, jurorswhereof@singles50s.com>)
Text 8: Replica watches, bags, pens (<Exquisite Replica, ctoni@altavista.nl>)
Text 9: Human Growth Hormone (4ever Young, <tduluna@smapxsmap.net>)

	Text 10		**Text 11**		**Text 12**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	124	119.6543	53	53.2943	58	58.7105
2	43	45.2920	27	26.8886	20	20.1095
3	20	25.3445	10	8.4085	22	16.9035
4	12	7.4304	1	2.4086	3	5.0034
5	2	3.2788			1	3.2732
∑	201		91		104	
	a = 0.5187, b = 0.2702 X^2 = 4.71, FG = 2 P = 0.09		a = 0.5350, b = 0.0570 X^2 = 1.13, FG = 1 P = 0.29		a = 0.5718, b = 0.4010 X^2 = 3.93, FG = 2 P = 0.14	

Text 10: I or Monroeville (“Belinda Worley”, <mattawan6@junkmail.com>)
Text 11: No an kaleva (“Alejandro Dudley”, <nsautrain@inststanthosting.net>)
Text 12: Most effective weight loss (Get Hoodia, <rvaldez@glay.org>)

	Text 13		**Text 14**		**Text 15**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	66	63.5279	65	64.4249	196	193.9624
2	29	29.2919	27	27.1790	68	67.5309
3	11	15.1620	9	10.0230	27	31.3846
4	9	7.0182	3	2.6160|	9	8.1984
5			1	0.7571|	3	2.3016|
6					0	0.4921|
7					1	0.1299|
∑	115		105		304	
	a = 0.5935, b = 0.2230 X^2 = 1.80, FG = 1 P = 0.1796		a = 0.4885, b = 0.1363 X^2 = 0.23, FG = 1 P = 0.6336		a = 0.4494, b = 0.2252 X^2 = 1.11, FG = 2 P = 0.5735	

Text 13: Urgent Stock Alert (“CNET:”, <lhdxpyhs@cafedelteatre.com>)
Text 14: Decrease fat reserves (Young Future, <lmarcy@marchmail.com>)

Text 15: Defence lawyer Bruce Cutler said the evidence would show the shot that killed actress Lana Clarkson was a “classic self-inflicted type of injury” (“Dicky Head”, <asgx@pacific.net.sg>)

	Text 16		**Text 17**		**Text 18**	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	56	54.0257	63	63.0420	66	62.7207
2	24	24.8741	27	26.9785	15	15.4333
3	14	17.2643	3	6.6062|	5	9.8081
4	9	6.1911	3	1.1802|	5	3.0379
5	3	3.6449	2	0.1932|		
∑	106		98		91	
	a = 0.6740, b = 0.3169 X^2 = 2.11, FG = 2 P = 0.35		a = 0.4412, b = 0.0300 C = 0.0000		a = 0.3722, b = 0.3388 X^2 = 3.81, FG = 1 P = 0.0510	

Text 16: Woking communities are diverse, reflecting the nation`s racial and ethnic demographics. (Tim, <vvs@canada.com>)
Text 17: Fw: Thank you for your recent refinance loan request, we are accepting your refinance loan request (“chance workin”, <zzfpokpp@rietinet.com>)
Text 18: loaded preserve (“Thornton”, <xeqcd@clinic.net>)

	Text 19		**Text 20**	
X	n_x	NP_x	n_x	NP_x
1	91	90.2736	102	102.4648
2	26	26.2003	28	27.9709
3	14	11.2154	21	17.5944
4*	0	3.3107	3	5.9699
∑	131		154	
	a = 0.3724, b = 0.2205 X^2 = 4.01, FG = 1 P = 0.05		a = 0.4074, b = 0.3300 X^2 = 2.14, FG = 1 P = 0.14	

* Bei Text 19 wurde eine zusätzliche Klasse viersilbiger Wörter mit null Beobachtungen eingefügt, um einen Freiheitsgrad zu gewinnen.

Text 19: Why VIAGRA? (Erectile Disfunction, <vebateman@pearl-online.de>)
Text 20: Do u remember (Maryjane Crawford, <marcusaneewtonasdy@ amaricanclinical.com>

7. Auswertung

Die Hirata-Poisson-Verteilung kann an alle von uns untersuchten Daten erfolgreich angepasst werden. Da Text 19 jedoch zu wenige Klassen beinhaltet, musste eine vierte, leere Klasse hinzugefügt werden. Das Ergebnis liegt zwar knapp unter dem Testkriterium, ist aber noch im tolerablen Bereich. In anderen Untersuchungen zu Wortlängen im Englischen hat sich von frühneuenglischer Zeit an bei Prosatexten die positive Singh-Poisson-Verteilung bewährt, die aber für die hier untersuchten Spam-E-Mails wesentlich schlechtere Ergebnisse erbrachte.

Man darf dies nicht überbewerten, da nur 20 Texte ausgewertet wurden. Es könnte sich aber durchaus um einen Effekt der Textsorte handeln.

8. Fazit

19 von 20 Texten ließen sich erfolgreich mit der Hirata-Poisson-Verteilung modellieren. Die meisten Anpassungen wiesen sogar sehr gute Testergebnisse auf. Fehlende Längenklassen wie bei Text 19 können jedoch bei so kurzen Texten gelegentlich zu Problemen führen. Bei der Interpretation des Ergebnisses ist zu beachten, dass es sich bei Spam-Mails um relativ kurze und heterogene Texte mit wenigen Längenklassen handelt, die einer Bearbeitung durch viele Autoren unterliegen. Unter diesen Umständen ist das Ergebnis in jedem Fall als gut einzustufen.

Im Rahmen der Untersuchung kann die Hypothese, dass Häufigkeitsverteilungen von Wortlängen in Texten einem gesetzmäßigen Verlauf folgen und mit einer Wahrscheinlichkeitsverteilung modellierbar sind, unterstützt werden.

Es gibt aufgrund der hier vorgelegten Ergebnisse keinen Grund, die Hypothese in Frage zu stellen, dass Wortlängen in englischen Texten einer Regularität folgen. Welche Rolle dabei die Textsorte auf die Wahl der Verteilung spielt, muss aber noch in weiteren Untersuchungen geklärt werden.

Literatur

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

Best, K.-H. (1998). Results and Perspectives of the Göttinger Project on Quantitative Linguistics. *Journal of Quantitative Linguistics 5, 155-162.*

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

Best, K.-H. (2005). *Linguistik in Kürze*, 3., überarbeitete Auflage, Göttingen. (Skript)

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

Cambridge English Pronouncing Dictionary (2006). Jones, Daniel (Hrsg.). Cambridge: Cambridge University Press.

Cassier, F.-U. (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, K.-H. (Hrsg.), *Häufigkeitsverteilungen in Texten*: *33-42*. Göttingen: Peust & Gutschmidt.

Hasse, A., Weinbrenner, M. (1997). Zur Häufigkeit von Wortlängen in englischen Texten. In: Best, K.-H. (Hrsg.): *Glottometrika 16,* 98-107. Trier: WVT.

Schmückle, B., Chi, T. (2004). Spam. Linguistische Untersuchung einer neuen Werbeform. In: Ruhnkehl, J., Schlobinski, P., Siever, T. (Hrsg.) Networx Nr. 39. http://www.mediensprache.net/networx/networx-39.pdf, Hannover. Zugriffsdatum 15.08.2007.

Stark, A.B. (2001). Die Verteilung von Wortlängen in schweizerdeutschen Privatbriefen. In: Best, K.-H. (Hrsg.), *Häufigkeitsverteilungen in Texten: 153-161*. Göttingen: Peust & Gutschmidt.

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

Wimmer, G., Altmann, G. (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 112-133.* Trier: WVT.

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

Glottometrics 17, 2008, 8-11

Das Fremdwortspektrum im Türkischen

Karl-Heinz Best, Göttingen

Abstract. In this paper, Altmann's (1993) model for rank orders will be tested as model for the diversification of loanwords in Türkish. The data are taken from Aktaş (2008).

Keywords: Turkish, loanwords

1. Fremdwörter im Türkischen

In letzter Zeit sind speziell die deutsch-türkischen Sprachbeziehungen mehrfach behandelt worden, und zwar aus zwei Perspektiven: 1. unter dem Gesichtspunkt, wann und wie viele Fremdwörter aus dem Türkischen ins Deutsche gelangt sind (Best 2001, 2005b, Körner 2004), 2. mit der Frage, wie viele deutsche Entlehnungen sich neben denen aus anderen Sprachen im Türkischen finden (Sağlam 2004). Für diesen zweiten Aspekt konnte gezeigt werden, dass die Häufigkeit, mit der Entlehnungen[1] aus verschiedenen Sprachen im Türkischen vertreten sind, einem Sprachgesetz unterliegen (Best 2005a). Eine neue Untersuchung von Aktaş (2008) gibt Gelegenheit, dieses Thema auf einer etwas aktualisierten und erweiterten Datenbasis noch einmal aufzugreifen.

Die Untersuchung von Aktaş (2008) stützt sich auf Daten, die aus dem Wörterbuch der türkischen Sprachgesellschaft (TDK) gewonnen wurden (*Türkçe Sözlük* 2005), das insgesamt 77407 Lemmata enthält; als Quelle wird die Internetadresse www.tdk.gov.tr, Stand Mai 2006, genannt.

2. Entlehnungen aus der Sicht der Quantitativen Linguistik

Abgesehen von den reinen Zahlenverhältnissen hat die Quantitative Linguistik zwei Perspektiven auf Entlehnungen. Ausgehend von der Annahme, dass Prozesse und Zustände in der Sprache Gesetzen unterliegen (Altmann 1985a, 7: „In der Sprache muß es ... Gesetze geben, wie überall in der Natur."), ist 1. die Frage zu stellen, ob die Übernahme von Fremdwörtern ins Türkische gemäß dem logistischen Gesetz erfolgt; diese Frage lässt sich noch nicht beantworten, da auch Aktaş (2008) keine Daten zum Verlauf der Entlehnungen angibt. 2. kann man jedoch untersuchen, ob die Anteile der verschiedenen Geber-Sprachen am türkischen Wortschatz dem sog. Diversifikationsgesetz (Altmann 1985, 2005) entspricht. Dieses Gesetz hat die formale bzw. funktional-semantische Differenzierung einer Einheit zum Gegenstand (Altmann 1985, 1996; Rothe 1991). Die verschiedenen Formen (= Diversifikation auf der Ausdrucksebene) bzw. funktional-semantischen Geltungen einer Entität erscheinen gemäß einem Sprachgesetz, das Altmann (1991) abgeleitet und begründet hat. Ein weiterer Typ von Diversifikation ist zu beobachten, wenn man untersucht, aus welchen Sprachen der Wortschatz einer Sprache oder eines Textes stammt (Rothe 1991: 30). In Best (2005b: 94) konnte bereits einmal für das Fremdwortspektrum des Türkischen gezeigt werden, dass die Verteilung von Entlehnungen in diese Sprache sich tatsächlich gesetzmäßig verhält. In dieser Untersuchung wurden die Geber-Sprachen der Fremdwörter in eine Rangfolge gebracht, bei der das Arabische als einflussreichste Sprache an erster Stelle steht, es folgen Französisch, Persisch, Grie-

[1] Es wird hier nicht zwischen Lehnwörtern und Fremdwörtern unterschieden.

chisch und weitere 20 Sprachen.

Für solche Rangordnungen hat Altmann (1993: 61f.) das Modell

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

abgeleitet, der auch dafür geeignet sein sollte, die Gesetzmäßigkeit der Entlehnungen ins Türkische zu erfassen. Er bewährt sich tatsächlich bei den Daten, die Aktaş (2008) bereitstellt. Die Ergebnisse finden sich in der folgenden Tabelle und Graphik. Laut Aktaş (2008: 74) hat das Wörterbuch seit seiner Ausgabe von 1998 von 60120 Stichwörtern auf 77407 zugenommen; der Fremdwortschatz stieg dabei von 13205 auf 14816 Einheiten.

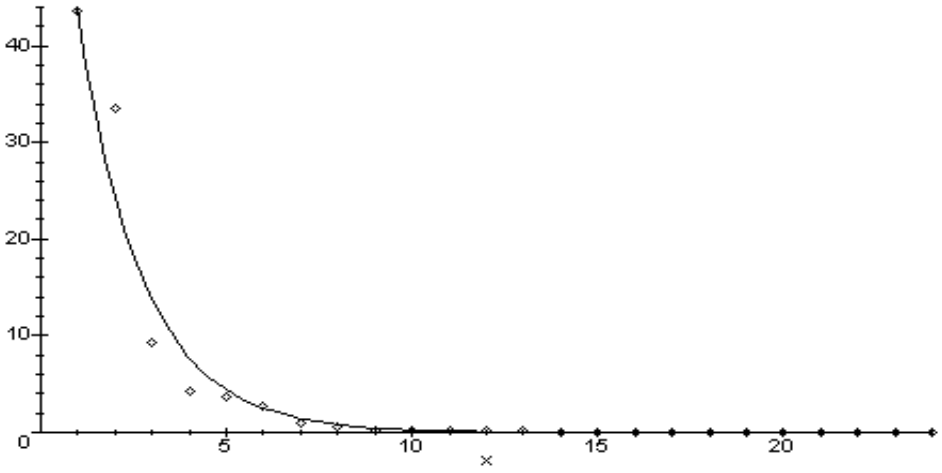
3. Modellierung der Entlehnungen ins Türkische

Die Anpassung von Altmanns Modell an die Entlehnungen ins Türkische ergab Resultate in Tabelle 1.

Tabelle 1
Entlehnungen ins Türkische 2006 (14816 Entlehnungen unter 77407 Stichwörtern)

Rang	Geber-Sprache	Entlehnungen absolut	Entlehnungen relativ	Entlehnungen berechnet
1	Arabisch	6463	43.62	43.62
2	Französisch	4974	33.57	24.41
3	Persisch	1374	9.27	13.75
4	Italienisch	632	4.27	7.80
5	Englisch	538	3.63	4.46
6	Griechisch	399	2.69	2.56
7	Latein	147	0.99	1.48
8	Deutsch	85	0.57	0.86
9	Russisch	40	0.27	0.50
10	Spanisch	36	0.24	0.30
11	Slawisch (ohne Russ.)	24	0.16	0.18
12	Armenisch	23	0.16	0.10
13	Ungarisch	19	0.13	0.06
14	Neugriechisch	14	0.09	0.04
15	Mongolisch	13	0.09	0.02
16	Hebräisch	9	0.06	0.01
17	Bulgarisch	8	0.05	0.01
18	Japanisch	7	0.05	0.01
19	Portugiesisch	4	0.03	0.00
20	Finnisch	2	0.01	0.00
21	Norwegisch	2	0.01	0.00
22	Albanisch	1	0.01	0.00
23	Koreanisch	1	0.01	0.00
24	Soghdisch	1	0.01	0.00
a = 109.5376		b = 60.4122	D = 0.9571	

Dabei sind *a* und *b* die Parameter des Modells, y_1 ist die empirische relative Häufigkeit der ersten Klasse. *D* ist der Determinationskoeffizient, der mit $D \geq 0.90$ eine sehr gute Übereinstimmung des Modells mit den beobachteten Daten anzeigt, wie dies auch hier der Fall ist und in Graphik 1 deutlich wird.



Graphik 1: Entlehnungen ins Türkische (Stand 2006)

4. Ergebnis und Perspektive

Laut Aktaş finden sich im türkischen Wortschatz 85 Entlehnungen aus dem Deutschen; sie machen 0.11% des gesamten erfassten Wortschatzes aus und 0.57% des Fremdwortschatzes (Aktaş 2008: 75). In der Dimension wird damit Stiberc (1999: 52) geringfügig übertroffen, die von „rund 70 Entlehnungen aus dem Deutschen“ spricht. Das Deutsche nimmt unter den Geber-Sprachen für das Türkische derzeit anscheinend Platz 8 ein; der Abstand zu den wichtigsten Geber-Sprachen *Arabisch*, *Französisch* und *Persisch* ist jedoch ganz erheblich.

Aus der Perspektive der Quantitativen Linguistik ist sehr wichtig, dass die Relationen zwischen den Wörtern unterschiedlicher Herkunft offenbar wie angenommen gesetzmäßig geregelt sind. Im Falle der Entlehnungen in englischen Pressetexten bewährte sich eher die negative Binomial-Poisson-Verteilung als geeignetes Modell (Best 2006: 89). Altmanns Modell für beliebige Rangfolgen hat sich jedoch bei den Fremdwörtern im Türkischen als sehr erfolgreich erwiesen; damit wird das Ergebnis der Untersuchung von (Best 2005a) bestätigt. Altmanns Vorschlag scheint den verschiedenen Modellen, die bisher für Diversifikationsphänomene bevorzugt wurden, zumindest bei den Entlehnungen deutlich überlegen zu sein. Für den Fall, dass man es als Wahrscheinlichkeitsverteilung verwendet, stellt y_1 die Normierungskonstante dar. Die Erfahrungen hierzu sind allerdings noch nicht besonders breit gestreut. Dieses Ergebnis bedarf daher weiterer Überprüfung, wie auch der Blick auf das Englische zeigt.

Die historischen Prozesse, die zu der heutigen Diversifikation des türkischen Wortschatzes geführt haben, sind einstweilen noch nicht so erfasst, dass man ihren gesetzmäßigen Verlauf testen könnte. Hierin liegt ein weiteres Desiderat der Forschung im Rahmen der Quantitativen Linguistik.

Literatur

Aktaş, Ayfer (2008). Aus dem Deutschen ins Türkische übernommene Wörter in türkischen Wörterbüchern – eine Bestandsaufnahme. *Muttersprache 118, 72-82.*

Altmann, Gabriel (1985). Semantische Diversifikation. *Folia Linguistica XIX, 177-200*.

Altmann, Gabriel (1985a). Sprachtheorie und mathematische Modelle. Christian-Albrechts-Universität Kiel, *SAIS Arbeitsberichte. H. 8: 1-13*.

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

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

Altmann, Gabriel (1996). Diversification processes of the word. In: Schmidt, Peter (Hrsg.), *Glottometrika 15: 102-111*. Trier: Wissenschaftlicher Verlag Trier.

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 646-658.* Berlin/ NewYork: de Gruyter.

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

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

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

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

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

Rothe, Ursula (1991). Diversification Processes in Grammar. An Introduction. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar: 3-32.* Hagen: Margit Rottmann Medienverlag.

Sağlam, Musa Yaşar (2004). Lehnwörter im Türkischen. Eine lexikologische Wortschatz-untersuchung. *Muttersprache 114, 115-122.*

Stiberc, Andrea (1999). *Sauerkraut, Weltschmerz, Kindergarten und Co. Deutsche Wörter in der Welt.* Freiburg/ Basel/ Wien: Herder.

Türkçe Sözlük. Hrsg. von Türk Dil Kurumu Lugat Kolu, TDK. 9. Auflage. Ankara 2005.

Verwendete Software

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

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

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

Glottometrics 17, 2008, 12-17

Pluralallomorphe in den Kurzgeschichten von Wolfdietrich Schnurre

Katharina Meuser, Jana Madlen Schütte, Sina Stremme[1]*,*

Universität Göttingen

Abstract. In this contribution the negative hypergeometric distribution has been fitted to the ranked distribution of plural morphemes in short stories of Wolfdietrich Schnurre. This paper brings a further corroboration of the hypothesis that the frequency distributions of entities contained in linguistic classes abide by laws.

Keywords: Plural, allomorph, German

1. Die Verteilung von Pluralallomorphen in deutschen Texten

Die vorliegende Untersuchung entstand im Rahmen des Göttinger *Projekts Quantitative Linguistik*, das sich der Erforschung von Sprachgesetzen widmet. Zu diesen Sprachgesetzen zählt das Diversifikationsgesetz. Es handelt sich dabei um ein Gesetz für die formale bzw. die funktional-semantische Differenzierung einer Einheit. Die unterschiedlichen Formen einer Einheit oder Sprachkategorie wie z.B. die Wortarten oder die Flexionsformen eines Substantivs oder Verbs folgen einer gesetzmäßigen Häufigkeitsverteilung. Dieses Sprachgesetz wurde von Altmann abgeleitet und begründet (Altmann 1991). Bisher gibt es eine Vielzahl von Überprüfungen, die das Diversifikationsgesetz stützen. Die Untersuchungen beschäftigen sich besonders mit der Verteilung von Wortarten. Ein weniger gut erforschtes Beispiel für die formale Diversifikation ist das Auftreten des Plurals bei deutschen Substantiven in Form von neun Allomorphen (Best 2006). Die Pluralallomorphe eignen sich besonders für eine Untersuchung, da es sich um ein differenzierbares Phänomen handelt und z.B. im Gegensatz zu einer Verteilung des Numerus oder der Person bis zu neun Klassen möglich sind.

Es liegt zu diesem Untersuchungskomplex bisher nur eine Arbeit von Brüers und Heeren vor, die das Häufigkeitsvorkommen der Pluralallomorphe in 25 Briefen Heinrich von Kleists ausgewertet haben. Die Autorinnen verwenden als Modell die 1-verschobene geometrische Verteilung, die sich deshalb besonders eignet, weil in manchen Briefen nur wenige verschiedene Pluralallomorphe auftraten. Brüers und Heeren kommen zu dem Schluss, dass sich die Pluralallomorphe von deutschen Substantiven in Kleists Briefen entsprechend einer Häufigkeitsverteilung verhalten und damit dem Diversifikationsgesetz entsprechen. Allerdings lasse sich das Ergebnis noch nicht verallgemeinern, da weitere Untersuchungen zu diesem Themenbereich fehlten (Brüers & Heeren 2004).

Um die von Brüers und Heeren erzielten Ergebnisse weiter zu überprüfen, wurden Kurzgeschichten von Wolfdietrich Schnurre für eine neue Untersuchung ausgewählt. Bei Kurzgeschichten handelt es sich um in sich abgeschlossene und überschaubare Texte, die meist kurz und stilistisch einheitlich sind. Daher ist weitgehend Homogenität gewährleistet (Best 2006).

[1] Address correspondence to: katharinameuser@gmx.de or Jana-Madlen.Schuette@t-online.de or sina_stremme@yahoo.de

Die Kurzgeschichten von Schnurre sind ausreichend lang, so dass beim Testen keine Probleme auftreten können, die aus dem zu geringen Umfang einzelner Texte resultieren, wie dies vereinzelt bei Brüers und Heeren der Fall war. Es wurden daher 21 Kurzgeschichten von Schnurre aus den Büchern „Als Vaters Bart noch rot war“ und „Als Vater sich den Bart abnahm“ ausgewählt. Wolfdietrich Schnurre (1920 – 1989) begann nach dem zweiten Weltkrieg mit seiner schriftstellerischen Tätigkeit und gehörte zu den Begründern der Gruppe 47. Sein Werk wurde mehrfach preisgekrönt. „Als Vaters Bart noch rot war“ ist sein bekanntestes Buch. 1995 erschien aus dem Nachlass der zweite Band „Als Vater sich den Bart abnahm“, der aus Geschichten besteht, die der Verleger 1958 bei der Veröffentlichung des ersten Bandes zu anstößig fand. In den beiden Büchern versetzt Schnurre den Leser mit seinen Vater-und-Sohn-Geschichten in das Berlin der zwanziger und dreißiger Jahre und erzählt von dem Alltag eines arbeitslosen Vaters, der sich mit seinem Sohn vor der Kulisse der drohenden NS-Herrschaft allein durchschlagen muss. Daher erscheinen gerade die Kurzgeschichten Schnurres besonders interessant und einer näheren Auseinandersetzung wert.

Bei den ausgezählten Pluralallomorphen handelt es sich um: {-e, -en, -er, -s, -n, -0 [= Nullallomorph], Umlaut, Umlaut + -e, Umlaut + -er}. Pluraliatantum wie z. B. Eltern, Geschwister, Leute, Kosten und Gemüse wurden bei der Auszählung nicht berücksichtigt.

2. Das angewendete Verteilungsmodell

Damit eine auswertbare Anzahl an Pluralallomorphen zur Verfügung stand, wurden aus dem Korpus der Kurzgeschichten die längeren Geschichten ausgewählt.

In den Kurzgeschichten wurden immer mindestens 6 Klassen von Pluralallomorphen beobachtet, so dass die negative hypergeometrische Verteilung mit drei Parametern angewandt werden konnte. Altmann (2005: 655) nennt diese Verteilung als eine von vielen Möglichkeiten, um Diversifikationsprozesse zu modellieren. Da keine Klasse $x = 0$ angesetzt wird, muss die negative hypergeometrische Verteilung in 1-verschobener Form verwendet werden:

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

Die Berechnungen wurden mit dem Altmann-Fitter (1997) durchgeführt. Bei einer Anpassung an die Daten wird die Verteilung als zufriedenstellend angesehen, wenn $P \geq 0{,}05$.

In den Tabellen bedeuten:

U = Umlaut
x = Rang der Pluralallomorphklasse
n_x = beobachtete Häufigkeit
NP_x = aufgrund der 1-verschobenen negativen hypergeometrischen Verteilung berechnete Häufigkeit
K, M, n = Parameter der 1-verschobenen negativen hypergeometrischen Verteilung
X^2 = Chiquadrat
P = Überschreitungswahrscheinlichkeit des Chiquadrats
FG = Freiheitsgrade
| = Diese Linie zeigt an, dass die entsprechenden Klassen bei der Berechnung zusammengefasst wurden.

3. Anpassung des Modells an die Briefdateien

Die Anpassung der 1-verschobenen hypergeometrischen Verteilung an die Daten der 21 Kurzgeschichten von Wolfdietrich Schnurre hat die folgenden Ergebnisse erbracht:

Text 1: Als Vater sich den Bart abnahm

Rang	*x*	n_x	*NPx*
1	-{n}	9	9.02
2	-{e}	9	8.45
3	-{en}	7	7.52
4	-{er}	6	6.45
5	-{s}	5	5.29
6	U + -{e}	5	4.06
7	-{0}	3	2.78
8	U + -{er}	1	1.43

K = 3.0206 *M* = 1.0644
n = 7 X^2 = 0.4845
FG = 4 *P* = 0.98

Text 2: Fritzchen

Rang	*x*	n_x	*NPx*
1	-{n}	28	25.80
2	-{e}	16	18.46
3	U + -{e}	14	14.33
4	-{0}	13	11.00
5	-{en}	8	7.87
6	-{er}	3	4.55

K = 2.4729 *M* = 0.8103
n = 5 X^2 = 1.4198
FG = 2 *P* = 0.49

Text 3: Das Zeichen

Rang	*x*	n_x	*NPx*
1	-{n}	17	16.53
2	U + -{e}	7	8.62
3	U + -{er}	7	6.29
4	-{en}	6	4.97
5	-{e}	4	4.04
6	-{0}	3	3.27
7	-{s}	2	2.54
8	-{er}	2	1.73

K = 1.9234 *M* = 0.5495
n = 7 X^2 = 0.7908
FG = 4 *P* = 0.94

Text 4: Herr Kellotat oder die Weite der Meere

Rang	*x*	n_x	*NPx*
1	-{e}	20	18.45
2	-{n}	14	10.33
3	-{en}	4	7.17
4	U + -{e}	4	5.13
5	-{0}	3	3.57
6	-{s}	2	2.26
7	U	1	1.10

K = 2.5791 *M* = 0.6467
n = 6 X^2 = 3.2129
FG = 3 *P* = 0.36

Text 5: Rückkehr ins Paradies

Rang	*x*	n_x	*NPx*
1	-{n}	11	10.47
2	-{e}	8	8.92
3	-{0}	7	7.27
4	-{en}	6	5.63
5	U + -{er}	5	4.03
6	U + -{e}	2	2.53
7	-{er}	1	1.15

K = 3.2354 *M* = 1.0246
n = 6 X^2 = 0.5167
FG = 3 *P* = 0.92

Text 6: Glück und Glas

Rang	*x*	n_x	*NPx*
1	-{n}	20	20.46
2	-{0}	7	6.51
3	-{e}	5	4.26
4	U + -{e}	3	3.25
5	-{en}	2	2.66
6	U	2	2.26
7	-{er}	2	1.94
8	-{s}	2	1.66

K = 1.3790 *M* = 0.3207
n = 7 X^2 = 0.4634
FG = 4 *P* = 0.98

Text 7: Laterne, Laterne

Rang	*x*	n_x	*NPx*
1	-{n}	21	22.22
2	-{e}	18	14.97
3	-{0}	10	10.02
4	U + -{e}	8	8.11
5	-{en}	5	5.76
6	-{er}	3	3.81
7	-{s}	2	2.20|
8	U + -{er}	2	0.91|

K = 3.1588 *M* = 0.8044
n = 7 X^2 = 1.3043
FG = 3 *P* = 0.73

Text 8: Abstecher ins Leben

Rang	*x*	n_x	*NPx*
1	-{n}	59	58.89
2	-{0}	33	31.30
3	-{e}	20	23.13
4	-{en}	18	18.54
5	U + -{er}	18	15.28
6	U + -{e}	12	12.60
7	U	10	10.08
8	-{er}	7	7.18

K = 1.8701 *M* = 0.5554
n = 7 X^2 = 1.0509
FG = 4 *P* = 0.90

Text 9: Erfahrungen mit Zwergen

Rang	*x*	n_x	*NPx*
1	-{e}	12	11.82
2	-{en}	9	9.67
3	-{n}	8	7.97
4	U + -{er}	8	6.44
5	-{0}	4	5.00
6	-{er}	3	3.64
7	-{s}	3	2.34
8	U	1	1.12

K = 3.0224 *M* = 0.9445
n = 7 X^2 = 0.9385
FG = 4 *P* = 0.92

Text 10: Aller Glanz für Willi			
Rang	*x*	n_x	*NPx*
1	-{n}	51	52.01
2	-{en}	25	24.84
3	-{0}	14	15.33
4	U + -{e}	13	9.81
5	-{er}	5	6.12
6	-{e}	3	3.54
7	U + -{er}	2	1.75|
8	U	1	0.60|
K = 3.3681 *M* = 0.5985			
n = 7 X^2 = 1.6414			
FG = 3 *P* = 0.65			

Text 11: Des Hasen Heimgang			
Rang	*x*	n_x	*NPx*
1	-{n}	20	18.83
2	-{en}	13	14.61
3	-{e}	10	11.19
4	U + -{e}	9	8.26
5	-{0}	8	5.76
6	U + -{er}	3	3.66
7	U	1	1.97|
8	-{er}	1	0.73|
K = 3.6517 *M* = 0.9631			
n = 7 X^2 = 1.6161			
FG = 3 *P* = 0.66			

Text 12: Veitel und seine Gäste			
Rang	*x*	n_x	*NPx*
1	-{n}	20	14.62
2	-{er}	7	7.20
3	-{0}	7	5.72
4	U + -{e}	6	5.23
5	-{en}	5	5.24
6	-{e}	4	5.88
7	-{s}	4	9.10
K = 1.0473 *M* = 0.4588			
n = 6 X^2 = 5.8504			
FG = 3 *P* = 0.12			

Text 13: Flieder			
Rang	*x*	n_x	*NPx*
1	-{n}	18	16.99
2	-{0}	9	11.17
3	U + -{e}	9	8.60
4	U + -{er}	7	6.80
5	-{en}	6	5.28
6	-{e}	5	3.85
7	-{s}	1	2.31
K = 2.3141 *M* = 0.7223			
n = 6 X^2 = 1.6865			
FG = 3 *P* = 0.64			

Text 14: Kalünz ist keine Insel			
Rang	*x*	n_x	*NPx*
1	-{n}	125	121.00
2	-{e}	80	85.12
3	-{en}	63	61.81
4	U + -{e}	43	44.26
5	-{0}	42	30.67
6	U +-{er}	13	20.21
7	-{er}	8	12.35
8	U	8	6.72
9	-{s}	4	3.86
K = 4.3207 *M* = 0.8932			
n = 9 X^2 = 9.0333			
FG = 5 *P* = 0.11			

Text 15: Die Leihgabe			
Rang	*x*	n_x	*NPx*
1	-{n}	10	11.11
2	-{e}	10	6.75
3	-{er}	4	4.99
4	U + -{er}	3	3.84
5	-{0}	2	2.96
6	U + -{e}	2	2.23
7	U	2	1.69
8	-{s}	1	1.03|
9	-{en}	1	0.50|
K = 2.6488 *M* = 0.6811			
n = 8 X^2 = 2.6335			
FG = 4 *P* = 0.62			

Text 16: Die Flucht nach Ägypten			
Rang	*x*	n_x	*NPx*
1	-{n}	39	38.11
2	-{e}	27	27.70
3	U + -{e}	20	20.88
4	-{0}	14	15.54
5	-{en}	12	11.17
6	-{er}	11	7.58
7	U	4	4.69
8	-{s}	1	2.45|
9	U + -{er}	1	0.88|
K = 3.6452 *M* = 0.8866			
n = 8 X^2 = 2.4646			
FG = 4 *P* = 0.65			

Text 17: Der Verrat			
Rang	*x*	n_x	*NPx*
1	-{e}	18	17.60
2	-{n}	13	14.16
3	U + -{e}	11	9.29
4	-{0}	4	5.13
5	-{en}	2	2.22|
6	U +-{er}	1	0.60|
K = 5.2129 *M* = 1.2766			
n = 5 X^2 = 0.6775			
FG = 1 *P* = 0.41			

Text 18: Wovon man lebt			
Rang	*x*	n_x	*NPx*
1	-{n}	18	17.93
2	-{e}	12	11.62
3	-{0}	8	9.09
4	U + -{e}	8	7.36
5	-{er}	6	5.85
6	U + -{er}	4	4.15
K = 2.0126 *M* = 0.6900			
n = 5 X^2 = 0.2080			
FG = 2 *P* = 0.90			

Text 19: Der Morgen der Welt			
Rang	*x*	n_x	*NPx*
1	-{n}	33	32.55
2	-{0}	18	19.24
3	U + -{e}	15	14.52
4	-{e}	13	11.60
5	-{er}	8	9.39
6	-{en}	7	7.48
7	U + -{er}	7	5.64
8	U	3	3.59
K = 2.1210 *M* = 0.6322			
n = 7 X^2 = 0.9302			
FG = 4 *P* = 0.92			

Text 20: Onkel Aluco, einige Vögel, die Zeit			
Rang	*x*	n_x	*NPx*
1	-{n}	63	59.03
2	-{0}	19	27.13
3	-{en}	18	18.64
4	U + -{e}	17	14.05
5	-{e}	14	10.93
6	U	9	8.51
7	-{er}	8	6.47
8	-{s}	2	4.59
9	U + -{er}	2	2.66
K = 2.1121 *M* = 0.4950			
n = 8 X^2 = 6.2153			
FG = 5 *P* = 0.29			

Text 21: Die Verbündeten			
Rang	*x*	n_x	*NPx*
1	-{e}	11	11.25
2	-{n}	11	7.92
3	-{0}	3	5.68
4	U + -{e}	3	3.85
5	U + -{er}	3	2.30|
6	-{en}	1	1.00|
K = 3.1277 *M* = 0.8802			
n = 5 X^2 = 2.7997			
FG = 1 *P* = 0.09			

Wie aus den Tabellen ersichtlich, bewegt sich *M* im Intervall <0.5, 1.28> und *K* im Intervall <1.05, 5.21>. Dies sind zwar schmale Intervalle, jedoch kann man zeigen, dass *M* mit wachsendem *K* selbst steigt. Die ungefähre Abhängigkeit lässt sich in Form $M = 0.2057K^{0.1987}$ ausdrücken, wobei der Determinationskoeffizient $R^2 = 0.71$, die t-Tests für Parameter und der F-Test aber hoch signifikant sind. Da *K* immer größer ist als *M*, kann man annehmen, dass bei Anwachsen dieser beiden Parameter diese Diversifikationsart gegen die Binomialverteilung konvergieren würde; dies wäre der Fall, wenn die Diversifikation mehr Klassen umfasste oder besser ausgeprägt wäre, mit $K \to \infty$, $M \to \infty$, und $M/K \to p$. Zu diesem Zweck müssten weitere Texte von verschiedenen Schriftstellern oder aus verschiedenen Genres analysiert werden.

Abschließend werden zwei Anpassungen zur Veranschaulichung graphisch dargestellt:

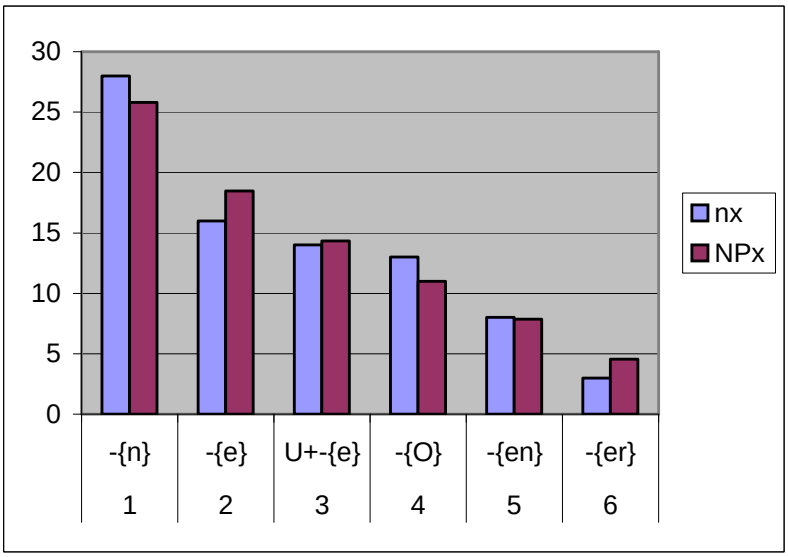


Abb. 1. Beispielgraphik zu Kurzgeschichte 2

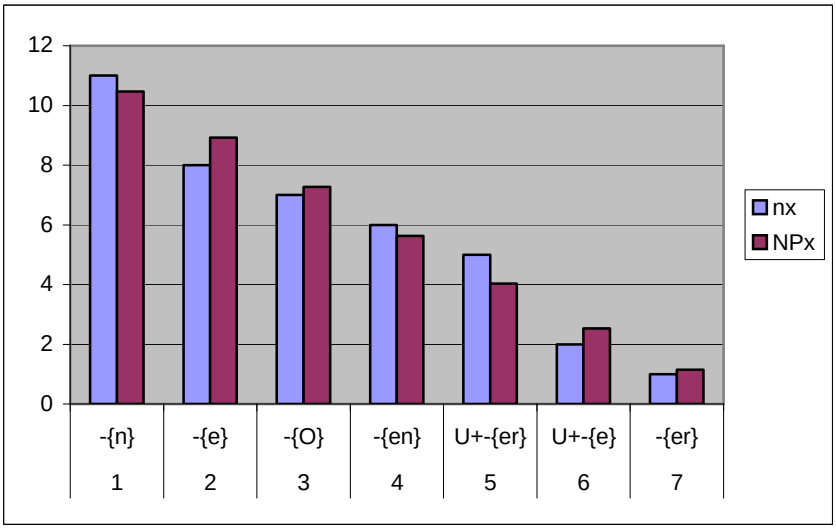


Abb. 2. Beispielgraphik zu Kurzgeschichte 5

4. Zusammenfassung

Die vorliegende Untersuchung hat ergeben, dass sich die Verteilung der Pluralallomorphe in den ausgezählten Kurzgeschichten mit der negativen hypergeometrischen Verteilung modellieren lässt. Somit stellt dieses Ergebnis eine Bestärkung der Hypothese dar, dass sprachliche Phänomene Gesetzen unterliegen. Speziell für das Diversifikationsgesetz und seine Anwendung auf die Pluralallomorphe sind jedoch weitere Untersuchungen angebracht.

Literatur

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

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 646-658.* Berlin/ N.Y.: de Gruyter.

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

Brüers, Nina, & Heeren, Anne (2004). Pluralallomorphe in Briefen Heinrich von Kleists. *Glottometrics 7, 85 – 90.*

Schnurre, Wolfdietrich (2000). *Als Vaters Bart noch rot war. Ein Roman in Geschichten.* München: Piper Verlag.

Schnurre, Wolfdietrich (1997). *Als Vater sich den Bart abnahm. Erzählungen.* München: Piper Verlag.

Software

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

Glottometrics 17, 2008, 18-42

Word frequency and arc length

Ioan-Iovitz Popescu, Bucharest[1]
Ján Mačutek, Bratislava
Gabriel Altmann, Lüdenscheid

Abstract. In the present article we try to exploit the arc length of the rank-frequency distribution of words for characterization of texts and typological purposes. Several indicators and functions are proposed and tested on 100 texts in 20 languages.

Keywords: word frequency, typology, arc length, rank-frequency distribution

1. The problem

A number of means have been used to characterize word frequency distributions and, using these distributions, a writer, genre or language. First, after almost a century of development the distribution itself acquired three basically different forms: (a) Zipf´s (zeta) distribution, developed further by Mandelbrot and many other researchers through generalization and modification of Zipf´s formula, (b) the negative hypergeometric distribution yielding excellent fits for rank-frequency data even in music (cf. Köhler, Martináková-Rendeková 1998; Popescu et al. 2008), and (c) the conception of text as a stratified entity leading to a superposition of exponentials (cf. Popescu, Altmann, Köhler 2008c). Second, different formulas and indicators of type-token ratios have been employed to trace the flow of information in text. Their disadvantage is the great confidence interval and in most cases a dependence on text length. And third, indicators characterizing different aspects of text have been used, such as vocabulary richness, thematic concentration, autosemantic compactness, entropy, synthetism etc. (cf. Popescu et al. 2008).

In the present article we try to test the possibility of using the arc length of rank-frequency distributions for text characterization and typological purposes. We start from the fact that arc length can be computed both empirically from the given data decreasing in rank-frequency and from the theoretical probability mass/density function expressing the course of data. Evidently, it can be assumed that long-tail texts have smaller arc-lengths because many words situated on the tail have very small differences in frequency, while those with short tails have greater differences and yield greater distances. It has been shown that texts written in highly synthetic languages have longer tails because not all forms of a word are repeated, many of which remain to be hapax legomena (Popescu, Altman 2008a,b). Hence, considering word forms, the ratio of arc length and vocabulary of the text in strongly synthetic languages must differ from that in strongly analytic languages. Let us first try to demonstrate this hypothesis.

Let arc length of a discrete probability mass function be defined as

$$(1) \qquad L = \sum_{r=1}^{V-1} \{[f(r) - f(r+1)]^2 + 1\}^{1/2}$$

[1] Address correspondence to: iovitzu@gmail.com

i.e. as the sum of Euclidean distances between two neighbouring frequencies, r being the rank, $f(r)$ the frequency at rank r. One segment of the arc is presented in Figure 1.

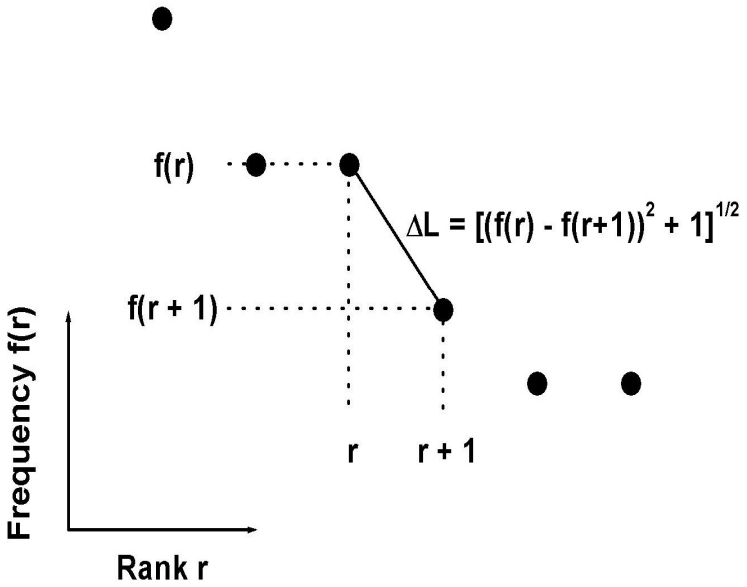


Figure 1. Computation of the length of one arc segment

For the sake of illustration we compute the arc length in a fictitious example. Let

r	$f(r)$
1	20
2	14
3	9
4	3
5	1
6	1

then

$$L = [(20-14)^2 + 1]^{1/2} + [(14-9)^2 + 1]^{1/2} + [(9-3)^2 + 1]^{1/2} + [(3-1)^2 + 1]^{1/2} + [(1-1)^2 + 1]^{1/2} = 20.50.$$

The shortest arc length is represented by the length of the straight line between $(1,f_1)$ and $(r_{max},1) = (V,1)$, the longest can be attained if all but the first rank have frequencies 1, i.e.

$$L_{\min} = [(V-1)^2 + (f_1 - 1)^2]^{1/2}$$

and

$$L_{\max} = [(f_1 - 1)^2 + 1)]^{1/2} + V - 2.$$

In particular, if all words have the same frequency $f(r) = 1$, then $L_{min} = L_{max} = V\text{-}1$. These cases are not met in sufficiently long texts.[2]

[2] An approximation to the arc length in terms of the highest frequency f_1, inventory V and the h-point can be computed as

$$L_{approx} = [(f_1 - h)^2 + (h - 1)^2]^{1/2} + [(V - h)^2 + (h - 1)^2]^{1/2}$$

The values provided by this expression are lower by about one percent than the observed ones.

Of course, in an absolute sense arc length depends on the vocabulary V of the text and the first frequency f_1 , hence individual differences can be tracked down.

In order to set up a comparable indicator one can go several ways. (1) To use the ratio

$$(2) \qquad B_1 = \frac{L}{L_{\max}}$$

or (2) to perform the normalization in the following way

$$(3) \qquad B_2 = \frac{L - L_{\min}}{L_{\max} - L_{\min}},$$

but since L depends on V and f_1, one can set up the indices

$$(4) \qquad B_3 = \frac{V - 1}{L}$$

$$(5) \qquad B_4 = \frac{f_1 - 1}{L}$$

or, finally, one can study the relationships

$$(6) \qquad L = f(f_1)$$

$$(7) \qquad B_3 = f(N).$$

In all cases we obtain very pronounced results both for individual texts and for individual languages.

2. The indicators

Let us begin with the indicators B_i (i = 1,2,3,4) whose values for 100 texts in 20 languages are presented in Table 1 (texts taken from Popescu et al. 2008, where the origin of texts is described).

Table 1
Indicators B_i of 100 texts in 20 languages
(B = Bulgarian, Cz = Czech, E = English, G = German, H = Hungarian, Hw = Hawaiian, I = Italian, In = Indonesian, Kn = Kannada, Lk = Lakota, Lt = Latin, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Ru = Russian, Sl = Slovenian, Sm = Samoan, T = Tagalog)

ID	*Text*	*V*	*f(1)*	*L*	B_1	B_2	B_3	B_4
B 01	Boris 2 (Letter)	400	40	428.45	0.980	0.763	0.931	0.091
B 02	Ceneva1 (Letter)	201	13	205.38	0.973	0.470	0.974	0.058
B 03	Ceneva 2 (Letter)	285	15	289.80	0.976	0.430	0.980	0.048
B 04	Janko1 (Letter)	286	21	297.03	0.977	0.618	0.959	0.067

B 05	Janko 3 (Letter)	238	19	247.30	0.974	0.588	0.958	0.073
Cz 01	Hrabal 310, Expozé panu ministru informací	638	58	684.17	0.987	0.835	0.931	0.083
Cz 02	Hrabal 315, Lednová povídka	543	56	586.22	0.984	0.809	0.925	0.094
Cz 03	Hrabal 316, Únorová povídka	1274	182	1432.06	0.986	0.875	0.889	0.126
Cz 04	Hrabal 319, Blitzkrieg	323	27	341.99	0.986	0.790	0.942	0.076
Cz 05	Hrabal 323, Protokol	556	84	626.98	0.984	0.868	0.885	0.132
E 01	Jimmy Carter, Nobel lecture (Peace 2002)	939	126	1042.85	0.982	0.834	0.899	0.120
E 02	Toni Morrison, Nobel lecture (Literature 1993)	1017	168	1157.22	0.979	0.837	0.878	0.144
E 03	George C. Marshall, Nobel lecture (Peace 1953)	1001	229	1204.91	0.982	0.890	0.830	0.189
E 04	James M. Buchanan Jr., Nobel lecture (Economy 1986)	1232	366	1567.31	0.983	0.911	0.785	0.233
E 05	Saul Bellow, Nobel lecture (Literature 1976)	1495	297	1760.86	0.984	0.894	0.848	0.168
E 07	Sinclair Lewis, Nobel lecture (Literature 1930)	1597	237	1800.70	0.983	0.861	0.886	0.131
E 13	Richard Feynman, Nobel lecture (Physics 1965)	1659	780	2388.47	0.980	0.921	0.694	0.326
G 05	Goethe - Der Gott und die Bajadere	332	30	351.41	0.979	0.716	0.942	0.083
G 09	Goethe - Elegie 19	379	30	398.43	0.981	0.718	0.949	0.073
G 10	Goethe - Elegie 13	301	18	309.84	0.980	0.602	0.968	0.055
G 11	Goethe - Elegie 15	297	18	306.80	0.983	0.664	0.965	0.055
G 12	Goethe - Elegie 2	169	14	175.44	0.974	0.601	0.958	0.074
G 14	Goethe - Elegie 5	129	10	132.54	0.974	0.546	0.966	0.068
G 17	Goethe - Der Erlkönig	124	11	127.96	0.969	0.527	0.961	0.078
H 01	Orbán Viktor beszéde az Astoriánál	1079	225	1288.83	0.991	0.939	0.836	0.174
H 02	A nominalizmus forradalma	789	130	907.18	0.990	0.925	0.869	0.142
H 03	Népszavazás	291	48	332.44	0.989	0.915	0.872	0.141
H 04	Egyre több	609	76	674.06	0.988	0.885	0.902	0.111
H 05	Kunczekolbász	290	32	314.40	0.986	0.837	0.919	0.099
Hw 03	Moolelo, Kawelo, Mokuna I	521	277	764.27	0.961	0.851	0.680	0.361
Hw 04	Moolelo, Kawelo, Mokuna II	744	535	1229.31	0.963	0.871	0.604	0.434
Hw 05	Moolelo, Kawelo, Mokuna III	680	416	1047.48	0.958	0.847	0.648	0.396
Hw 06	Moolelo, Kawelo, Mokuna IV	1039	901	1876.68	0.969	0.893	0.553	0.480
I 01	Pellico, Le mie prigioni	3667	388	4007.01	0.989	0.877	0.915	0.097
I 02	Manzoni, I promessi sposi	2203	257	2426.40	0.988	0.873	0.908	0.106
I 03	Leopardi, Canti	483	64	534.33	0.982	0.833	0.902	0.118

I 04	Deledda, Canne al vento	1237	118	1329.65	0.983	0.798	0.930	0.088
I 05	de Amicis, Il cuore	512	42	537.49	0.975	0.648	0.951	0.076
In 01	Assagaf-Ali Baba Jadi Asisten	221	16	228.49	0.976	0.590	0.963	0.066
In 02	BRI Siap Cetak Miliarder Dalam Dua Bulan	209	18	218.62	0.976	0.647	0.951	0.078
In 03	Pengurus PSM Terbelah	194	14	199.85	0.975	0.553	0.966	0.065
In 04	Pemerintah Andalkan Hujan	213	11	217.37	0.983	0.583	0.975	0.046
In 05	Fajar	188	16	195.65	0.973	0.599	0.956	0.077
Kn 003	Pradhana Gurudhat	1833	74	1891.11	0.993	0.817	0.969	0.039
Kn 004	Pradhana Gurudhat	720	23	733.26	0.991	0.673	0.981	0.030
Kn 005	T.R.Nagappa	2477	101	2558.43	0.994	0.829	0.968	0.039
Kn 006	T.R.Nagappa	2433	74	2481.41	0.991	0.681	0.980	0.029
Kn 011	D.N.S.Murthy	2516	63	2557.69	0.993	0.696	0.983	0.024
Lk 01	The fly on the window	174	20	184.77	0.967	0.632	0.936	0.103
Lk 02	Iktomi meets the prairie chicken	479	124	579.97	0.967	0.812	0.824	0.212
Lk 03	Iktomi meets two women	272	62	317.63	0.960	0.749	0.853	0.192
Lk 04	Bean, grass, and fire	116	18	125.56	0.958	0.630	0.916	0.135
Lt 01	Vergil, Georgicon liber primus	2211	133	2328.00	0.994	0.898	0.949	0.057
Lt 02	Apuleius, Fables, Book 1	2334	190	2502.00	0.992	0.895	0.932	0.076
Lt 03	Ovidius, Ars amatoria, liber primus	2703	103	2783.00	0.993	0.798	0.971	0.037
Lt 04	Cicero, Post reditum in senatu oratio	1910	99	1983.00	0.989	0.757	0.963	0.049
Lt 05	Martialis, Epigrammata	909	33	930.00	0.990	0.704	0.976	0.034
Lt 06	Horatius, Sermones.Liber 1, Sermo 1	609	19	621.00	0.994	0.760	0.979	0.029
M 01	Maori Nga Mahi	398	152	526.92	0.963	0.836	0.753	0.287
M 02	Ko Te Paamu	277	127	386.01	0.963	0.846	0.715	0.326
M 03	A Tawhaki	277	128	384.62	0.957	0.823	0.718	0.330
M 04	Ka Pu Te Ruha	326	137	444.29	0.966	0.854	0.732	0.306
M 05	Ka Kimi A Maui	514	234	715.18	0.960	0.836	0.717	0.326
Mq 01	Story Kopuhoroto'e	289	247	506.98	0.951	0.831	0.568	0.485
Mq 02	Ka´akai o Te Henua ´Enana	150	42	178.59	0.945	0.698	0.834	0.230
Mq 03	Te Hakamanu	301	218	500.37	0.970	0.893	0.600	0.434
Mr 001	Prof. B.P.Joshi, Nisar Sheti	1555	75	1612.43	0.991	0.795	0.964	0.046
Mr 018	V.L.Pandy, Thumcha chehara thumche yaktimatv	1788	126	1890.34	0.989	0.827	0.945	0.066
Mr 026	Kanchan Ganekar, Nath ha majha	2038	84	2098.93	0.991	0.750	0.970	0.040
Mr 027	Prof.Sarangar, Rashtriy Uthpann	1400	92	1467.65	0.986	0.755	0.953	0.062
Mr 288	Madhav Gadkari, Chaupher	2079	84	2141.01	0.991	0.764	0.971	0.039
R 01	Eminescu, Luceafarul	843	62	886.35	0.983	0.729	0.950	0.069

R 02	Eminescu, Scrisoarea III	1179	110	1269.07	0.987	0.836	0.928	0.086
R 03	Eminescu, Scrisoarea IV	719	65	770.20	0.986	0.820	0.932	0.083
R 04	Eminescu, Scrisoarea I	729	49	764.36	0.986	0.766	0.952	0.063
R 05	Eminescu, Scrisoarea V	567	46	599.19	0.982	0.744	0.945	0.075
R 06	Eminescu, Scrisoarea II	432	30	451.75	0.984	0.731	0.954	0.064
Rt 01	Kauraka Kauraka, Akamaramaanga	223	111	315.91	0.954	0.819	0.703	0.348
Rt 02	Tepania Puroku, Ko Paraka e te Kehe	214	69	264.75	0.946	0.730	0.805	0.257
Rt 03	Herekaiura Atama, Ko Tamaro e ana uhi	207	66	255.86	0.948	0.738	0.805	0.254
Rt 04	Temu Piniata, Te toa ko Teikapongi	181	49	215.58	0.950	0.719	0.835	0.223
Rt 05	Kaimaria Nikoro, Te toa ko Herehuaroa e Araitetonga	197	74	250.69	0.935	0.706	0.782	0.291
Ru 01	Tolstoy, Metel’	422	31	441.04	0.980	0.679	0.955	0.068
Ru 02	Dostoevskij, Prestuplenie i nakazanie (p. I, ch. 1)	1240	138	1356.70	0.987	0.858	0.913	0.101
Ru 03	Pelevin, Buben verchnego mira	1792	144	1909.09	0.988	0.825	0.938	0.075
Ru 04	Turgenev, Bežin lug	2536	228	2731.76	0.989	0.865	0.928	0.083
Ru 05	Gogol, Portret	6073	701	6722.04	0.993	0.926	0.903	0.104
Sl 01	Jurčič, Sosedov sin	457	47	493.72	0.985	0.829	0.924	0.093
Sl 02	Levstik, Zveženj	603	66	651.09	0.978	0.753	0.925	0.100
Sl 03	Grum, Vrata	907	102	990.94	0.985	0.840	0.914	0.102
Sl 04	Cankar, V temi	1102	328	1404.13	0.984	0.918	0.784	0.233
Sl 05	Kočevar, Grof in menih	2223	193	2385.35	0.989	0.849	0.932	0.080
Sm 01	O le tala ia Sina ma lana tuna	267	159	403.17	0.953	0.825	0.660	0.392
Sm 02	O le mea na maua ai le ava	222	103	303.92	0.944	0.770	0.727	0.336
Sm 03	O upu faifai ma le gaoi	140	45	168.39	0.925	0.624	0.825	0.261
Sm 04	O le faalemigao	153	78	214.17	0.939	0.760	0.710	0.360
Sm 05	Ataliitaga	124	39	149.49	0.934	0.664	0.823	0.254
T 01	Hernandez, Magpinsan	611	89	680.99	0.977	0.802	0.896	0.129
T 02	Hernandez, Limang Alas	720	107	807.46	0.980	0.830	0.890	0.131
T 03	Rosales, Kristal Na Tubig	645	128	748.50	0.972	0.811	0.860	0.170

If we look at the graphs of these indicators, we can see that some of them are not very discriminative, e.g. B_1. As can be seen in Figure 2, it is in all cases greater than 0.9. It can, of course, be presented in such a way that it shows the differences which clearly discriminate individual languages, but the small variation is not very helpful. One can clearly see the decreasing synthetism from left to right, surely corresponding with the character of these languages.

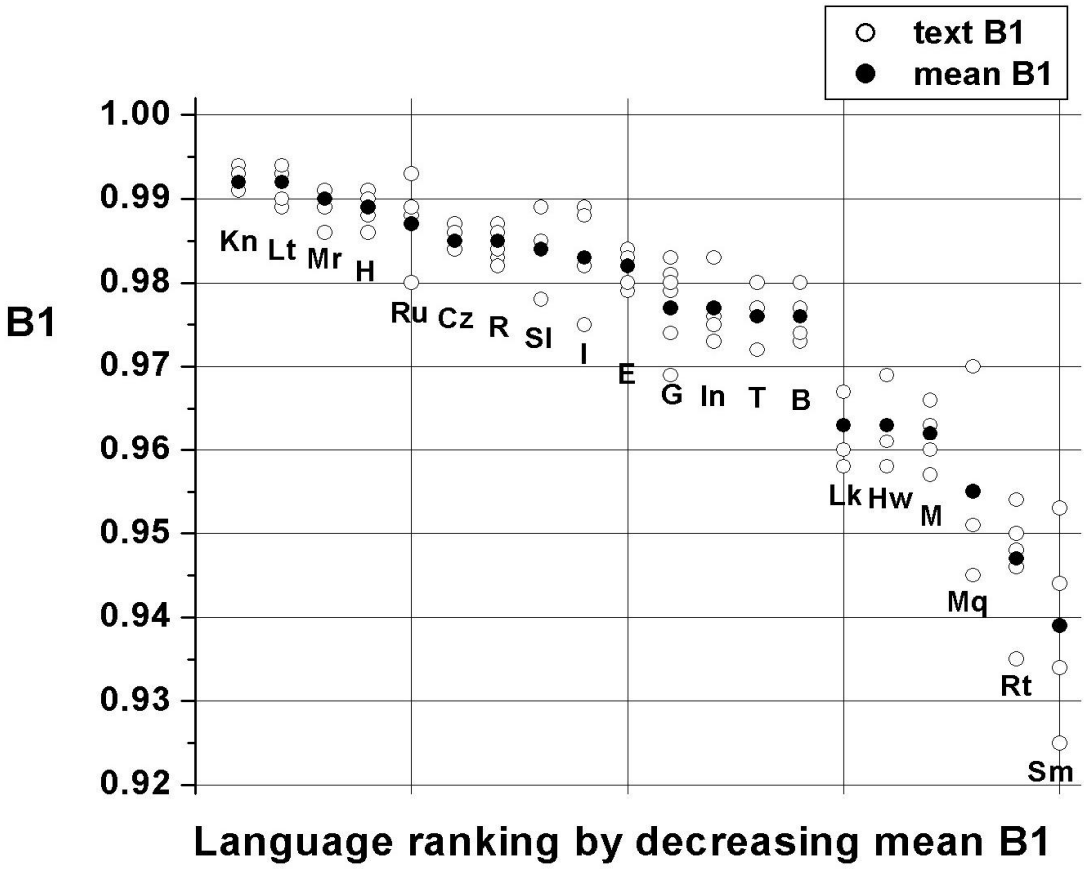


Figure 2. Indicator B_1 for 20 languages

The individual average values are presented in Table 2.

Table 2
Means of B_1 for individual languages

1	Kannada	0.992	11	German	0.977
2	Latin	0.992	12	Indonesian	0.977
3	Marathi	0.990	13	Tagalog	0.976
4	Hungarian	0.989	14	Bulgarian	0.976
5	Russian	0.987	15	Lakota	0.963
6	Czech	0.985	16	Hawaiian	0.963
7	Romanian	0.985	17	Maori	0.962
8	Slovenian	0.984	18	Marquesan	0.955
9	Italian	0.983	19	Rarotongan	0.947
10	English	0.982	20	Samoan	0.939

The indicator B_2 is normalized and displays a great variation. It is not correlated with other indicators, hence it can be used for characterization. Its disadvantage is not its computation but its statistical properties. In any case, it is theoretically positioned in the interval <0,1>. Though it perhaps never attains its extreme points, it discriminates the languages quite well. Its value for individual languages is shown in Figure 3 and Table 3.

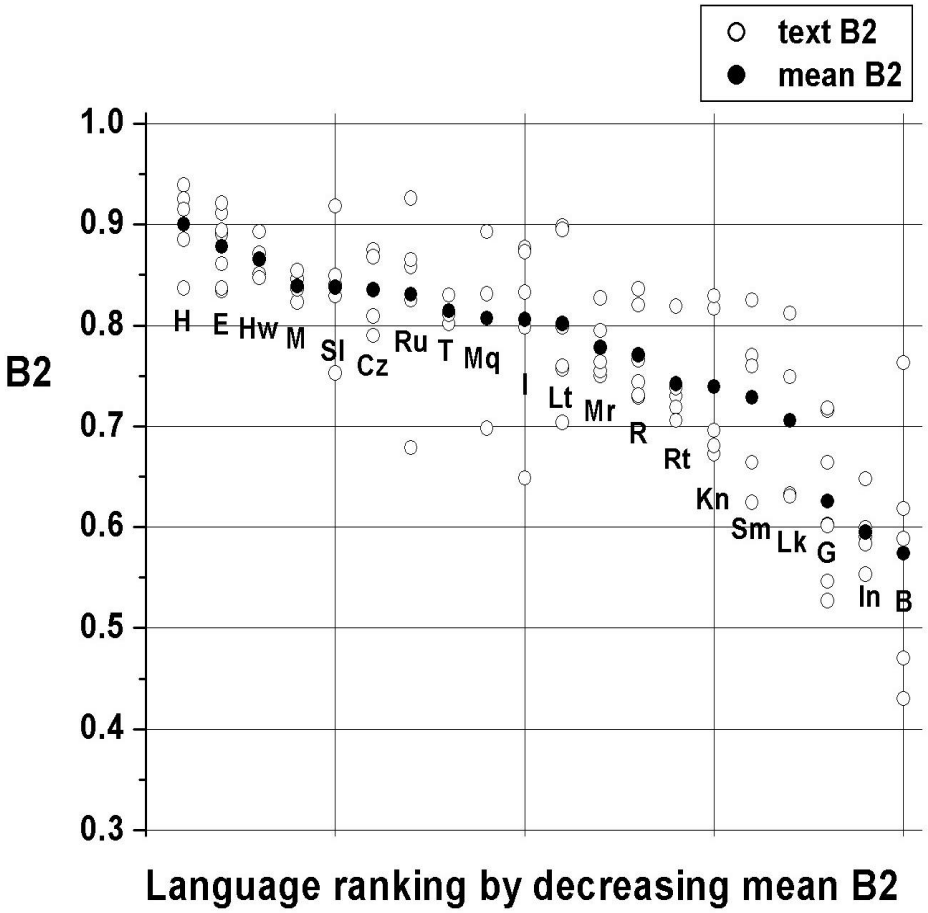


Figure 3. Indicator B_2 for 20 languages

Table 3
Means of B_2 for individual languages

1	Hungarian	0.900	11	Latin	0.802
2	English	0.878	12	Marathi	0.778
3	Hawaiian	0.866	13	Romanian	0.771
4	Maori	0.839	14	Rarotongan	0.742
5	Slovenian	0.838	15	Kannada	0.739
6	Czech	0.835	16	Samoan	0.729
7	Russian	0.831	17	Lakota	0.706
8	Tagalog	0.814	18	German	0.625
9	Marquesan	0.807	19	Indonesian	0.594
10	Italian	0.806	20	Bulgarian	0.574

Differences between the indicators B_1 and B_2 in two different texts can be tested as follows (we use the texts E12 and G01 as examples).

The statistics $\frac{B_{1,E12} - B_{1,G01}}{\sqrt{Var(B_{1,E12}) + Var(B_{1,G01})}}$ and $\frac{B_{2,E12} - B_{2,G01}}{\sqrt{Var(B_{2,E12}) + Var(B_{2,G01})}}$ have approximately the standard normal distribution, which means that the differences between the indicators are significant if their values exceed 1.96. Next, it holds

(8) $$Var(B_1) = Var\left(\frac{L}{L_{\max}}\right) = \frac{Var(L)}{L_{\max}^2}$$

and

(9) $$Var(B_2) = Var\left(\frac{L - L_{\min}}{L_{\max} - L_{\min}}\right) = \frac{Var(L)}{\left(L_{\max} - L_{\min}\right)^2}$$ (cf. Section 1 for $L_{\max}$ and $L_{\min}$).

The variances $Var(L_{E12}) = 646.67$ and $Var(L_{G01}) = 39.95$ were obtained by simulations. Hence we can evaluate

$$\frac{B_{1,E12} - B_{1,G01}}{\sqrt{\frac{Var(L_{E12})}{L_{\max,E12}^2} + \frac{Var(L_{G01})}{L_{\max,G01}^2}}} = \frac{0.0023}{0.0147} = 0.1564,$$

$$\frac{B_{2,E12} - B_{2,G01}}{\sqrt{\frac{Var(L_{E12})}{\left(L_{\max,E12} - L_{\min,E12}\right)^2} + \frac{Var(L_{G01})}{\left(L_{\max,G01} - L_{\min,G01}\right)^2}}} = \frac{0.0741}{0.0981} = 0.7556.$$

As can be seen, the differences are not significant in this case.

Indicator B_3 and its quasi-complement indicator B_4 (cf. $B_3 + B_4 \approx 1$) are simple ratios showing the relationship of arc length to the two extreme values f_1 and V. As shown in Figure 4, the lowest part of the relationship is occupied by strongly analytic languages while the upper part is rather mixed. Evidently, there are different boundary conditions in individual languages resulting in the given indicator value. In case of synthetism, perhaps different kinds of its building must be distinguished. Also the dispersions of the indicator in languages will be different, a circumstance depending most probably on the genre and author. The individual mean values of B_3 are shown in Table 4.

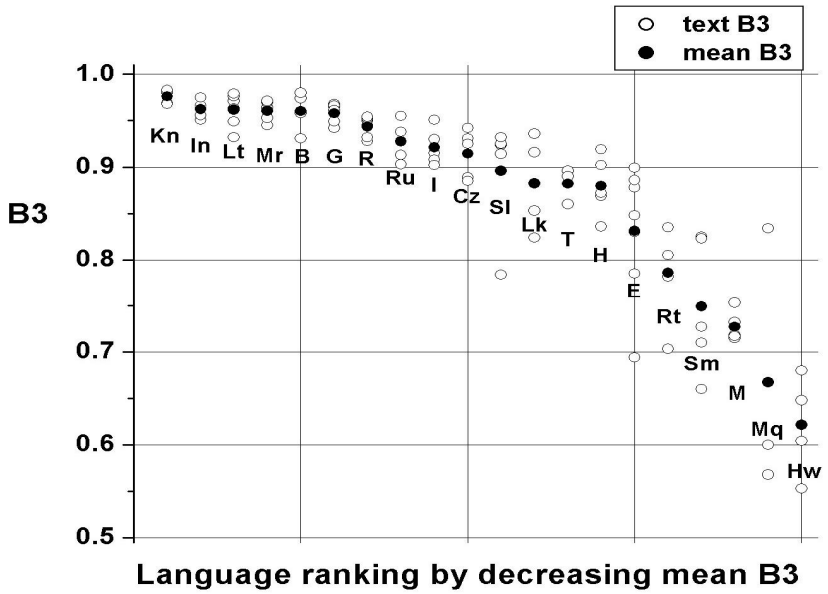


Figure 4. Indicator B_3 for 20 languages

Table 4
Means of B_3 for individual languages

1	Kannada	0.976	11	Slovenian	0.896
2	Indonesian	0.962	12	Lakota	0.882
3	Latin	0.962	13	Tagalog	0.882
4	Marathi	0.961	14	Hungarian	0.880
5	Bulgarian	0.960	15	English	0.831
6	German	0.958	16	Rarotongan	0.786
7	Romanian	0.944	17	Samoan	0.749
8	Russian	0.927	18	Maori	0.727
9	Italian	0.921	19	Marquesan	0.667
10	Czech	0.914	20	Hawaiian	0.621

Here, one cannot explain why e.g. Indonesian is in place 2 and Hungarian in place 14, but this is a problem of boundary conditions. This indicator could be correlated with Greenberg-Krupa and other indices (cf. Greenberg 1960; Krupa 1965) in order to search for boundary conditions, a task to be undertaken in further research. The values of B_4 are presented in Figure 5 and Table 5.

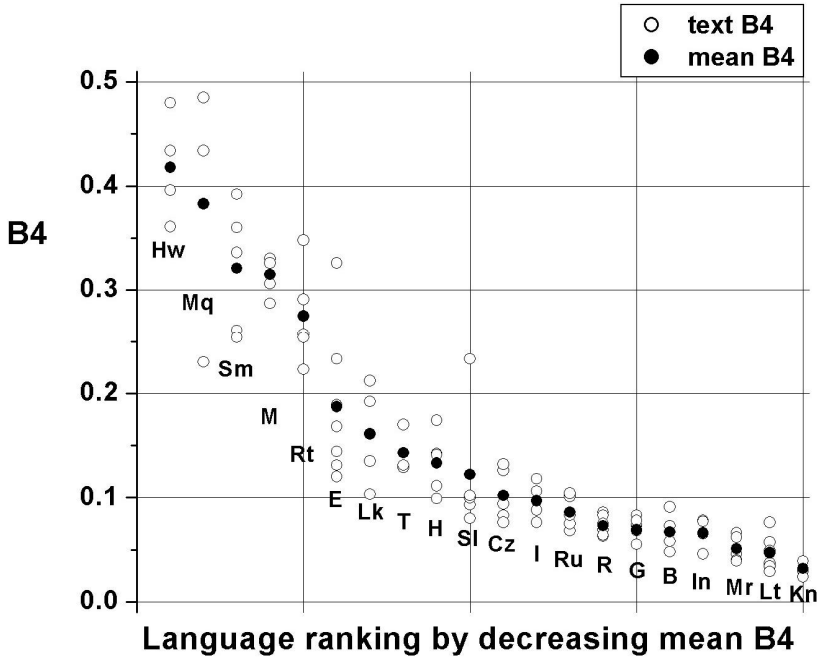


Figure 5. Indicator B_4 for 20 languages

Table 5
Means of B_4 for individual languages

1	Hawaiian	0.418	11	Czech	0.102
2	Marquesan	0.383	12	Italian	0.097
3	Samoan	0.321	13	Russian	0.086
4	Maori	0.315	14	Romanian	0.073
5	Rarotongan	0.275	15	German	0.069
6	English	0.187	16	Bulgarian	0.067

7	Lakota	0.161	17	Indonesian	0.066
8	Tagalog	0.143	18	Marathi	0.051
9	Hungarian	0.133	19	Latin	0.047
10	Slovenian	0.122	20	Kannada	0.032

The comparative survey of all languages is in Figure 6.

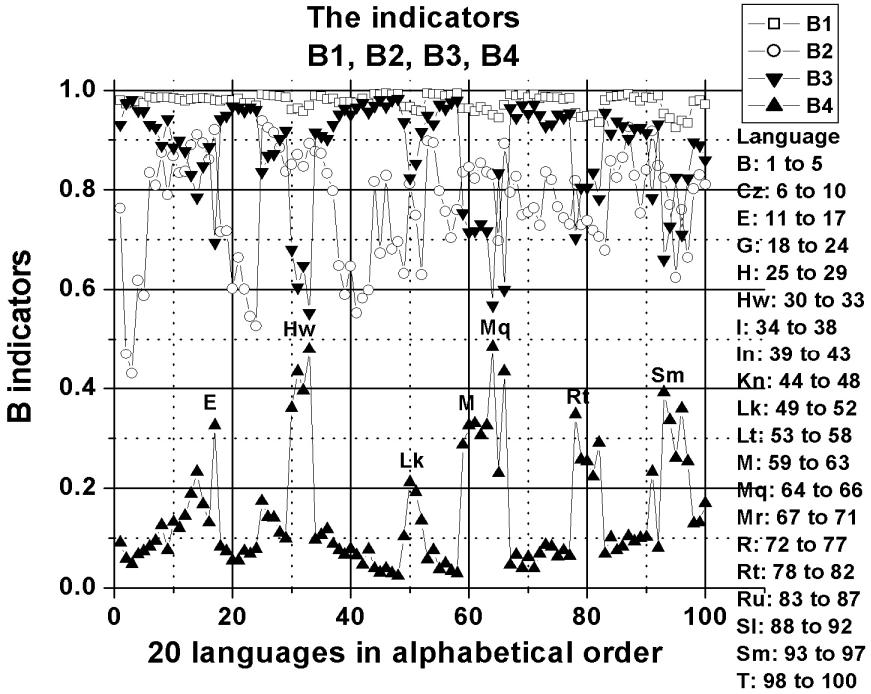


Figure 6. Comparison of all indicators

The complementarity of B_3 and B_4 can be better seen in Figure 7.

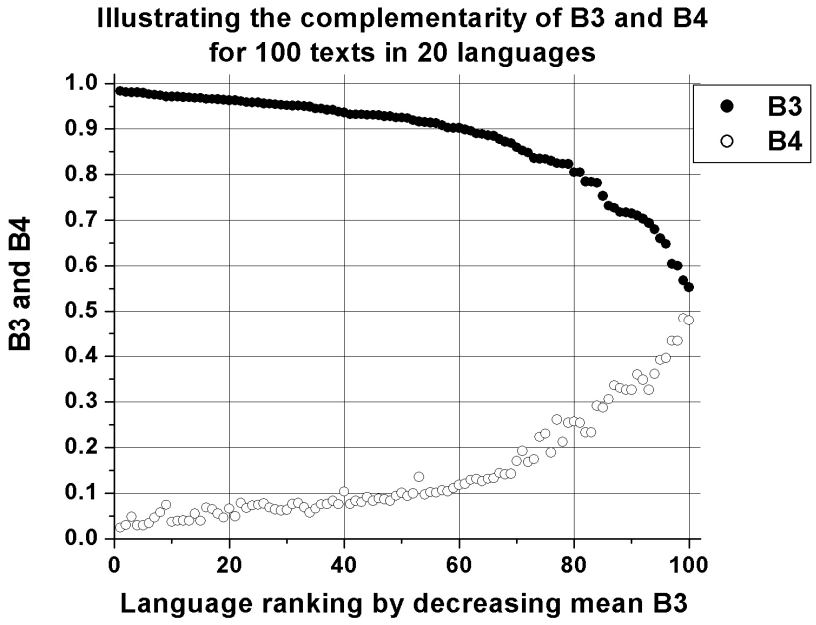


Figure 7. Complementarity of B_3 and B_4

3. Development of the arc

Because of the rigorous construction of texts, the arc exhibits a strong regularity along with text length N, therefore the individual indicators may change, too. Consider e.g. $B_3 = f(N)$. Since B_3 differs according to language, we must compare texts from one language only. To this end we present this indicator in 52 German texts of different length, as shown in Table 6. We took whole texts or sections of them. B_3, which is in the last column of the table, evidently decreases with increasing N. This trend can be lucidly shown in Figure 8. The trend can be captured by the power curve $B_3 = 1.1563N^{-0.0317}$ yielding a highly significant result (using the F-test) and a determination coefficient $R^2 = 0.73$ which is sufficient because we mixed texts of different kinds.

Thus when comparing two texts, their B_3-s must be normalized using the variances of this indicator. We suppose that for other languages the indicator will develop in the same way but with different parameters of the power curve. Thus languages can be compared using their theoretical B_3 curves.

Table 6
Dependence of B_3 on text length N

Text	*N*	*V*	*f(1)*	*L*	B_3
Schiller - Der Taucher	1095	530	83	598.77	0.883
Anonym - Fabel - Zaunbär	845	361	48	396.78	0.907
Krummacher - Das Krokodil	500	281	33	304.25	0.920
Anonym - Fabel - Mäuschen	545	269	32	289.67	0.925
Goethe - Der Gott und die Bajadere	559	332	30	351.41	0.942
Sachs - Das Kamel	545	326	30	346.82	0.937
Heine - Belsazar	263	169	17	178.72	0.940
Droste - Der Geierpfiff	965	509	39	534.55	0.950
Goethe - Elegie 19	653	379	30	398.43	0.949
Goethe - Elegie 13	480	301	18	309.84	0.968
Goethe - Elegie 15	468	297	18	306.80	0.965
Goethe - Elegie 2	251	169	14	175.44	0.958
Fontane - Gorm Grimme	460	253	19	262.25	0.961
Goethe - Elegie 5	184	129	10	132.54	0.966
Moericke - Peregrina	593	378	16	385.09	0.979
Lichtwer - Die Rehe	518	292	16	299.24	0.972
Goethe - Der Erlkönig	225	124	11	127.96	0.961
Heine - Vitzliputzli Präludium	356	227	15	234.23	0.965
Heine - Vitzliputzli I	986	561	37	585.28	0.957
Heine - Vitzliputzli II	683	411	35	436.46	0.939
Heine - Vitzliputzli III	715	421	28	438.45	0.958
Heine - Waldeinsamkeit	929	502	33	523.39	0.957
Heine - Spanische Atriden	1328	718	53	756.35	0.948
Heine - Prinzessin Sabbath	717	449	40	477.48	0.938
Heine - Disputation	2025	1024	85	1088.95	0.939

Schiller - Das Lied von der Glocke	2063	1029	97	1106.22	0.929
Kant - Kritik der reinen Vernunft - excerpt 01	4047	963	147	1078.03	0.892
Kant - Kritik der reinen Vernunft - excerpt 02	2326	681	100	758.85	0.896
Kant - Kritik der reinen Vernunft - excerpt 03	1630	512	81	575.28	0.888
Kant - Kritik der reinen Vernunft - excerpt 05	1096	374	53	412.00	0.905
Kant - Kritik der reinen Vernunft - excerpt 06	4412	1052	157	1177.79	0.892
Kant - Kritik der reinen Vernunft - excerpt 07	1649	570	69	621.21	0.916
Kant - Kritik der reinen Vernunft - excerpt 08	4515	1051	157	1175.11	0.894
Kant - Kritik der reinen Vernunft - excerpt 09	2909	1036	132	1146.45	0.903
Kant - Kritik der reinen Vernunft - excerpt 10	3253	841	143	956.11	0.879
Kant - Kritik der reinen Vernunft - excerpt 11	5490	1343	270	1579.28	0.850
Kant - Kritik der reinen Vernunft - excerpt 12	6869	1463	315	1736.34	0.842
Kant - Kritik der reinen Vernunft - excerpt 13	4043	1148	178	1296.99	0.884
Freud - Über Psychoanalyse - excerpt 1	3834	1483	126	1583.03	0.936
Freud - Über Psychoanalyse - excerpt 2	2617	1035	94	1196.08	0.864
Freud - Über Psychoanalyse - excerpt 3	3709	1354	147	1473.74	0.918
Freud - Über Psychoanalyse - excerpt 4	3012	1264	127	1368.00	0.923
Goethe - Novelle	7110	2469	276	2706.89	0.912
Goethe - Reineke Fuchs - excerpt 1-2	5486	1824	257	2048.33	0.890
Goethe - Reineke Fuchs - excerpt 1-2-3	9788	2614	454	3023.68	0.864
Goethe - Reineke Fuchs - excerpt 1-2-3-4	12656	3073	591	3612.01	0.850
Goethe - Reineke Fuchs - excerpt 5-6	6901	1939	329	2227.08	0.870
Goethe - Reineke Fuchs - excerpt 5-6-7	9493	2385	485	2823.88	0.844
Goethe - Reineke Fuchs - excerpt 5-6-7-8	12879	2951	656	3553.43	0.830
Goethe - Reineke Fuchs - excerpt 9-10	8426	2276	403	2637.29	0.863
Goethe - Reineke Fuchs - excerpt 10-11	8704	2413	406	2774.59	0.869
Goethe - Reineke Fuchs - excerpt 10-11-12	12335	3042	596	3589.93	0.847

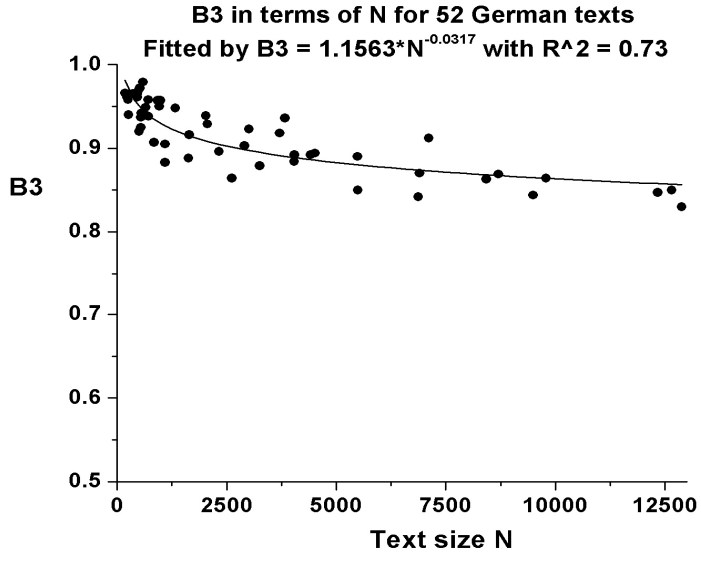


Figure 8. Dependence of B_3 on N

In the same way one may expect that L develops depending directly on f_1 – which itself depends on N according to the "type" of language – because the greater f_1, the longer must be the arc. As a matter of fact, for our German data the dependence is $L = 281.0546 + 5.7048 f_1$ with $R^2 = 0.95$, and $L = 32.8893 f_1^{0.7302}$ with $R^2 = 0.96$, see Figure 9. But again, for different languages the parameters may be different.

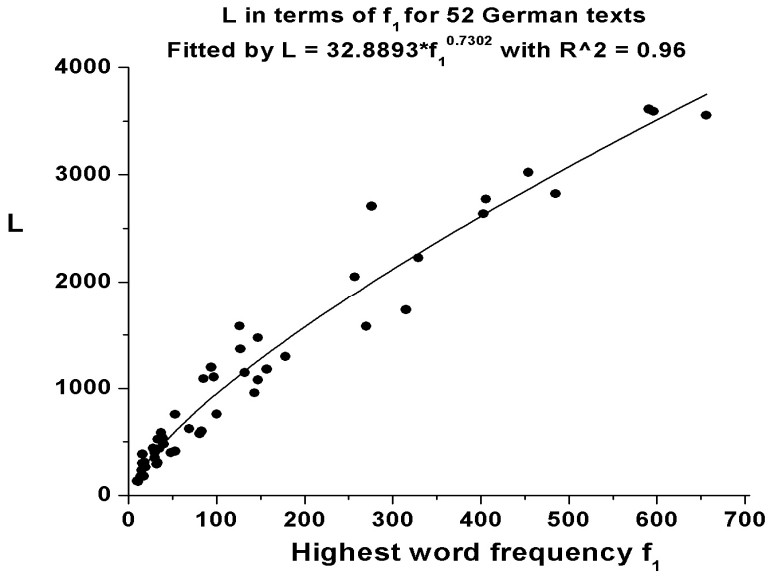


Figure 9. Dependence of L on f_1 for 52 German texts

In typologically different languages the $L = f(f_1)$ curves will be different. As can be seen in Figure 10, Hawaiian, English, German and Latin differ drastically, as can be expected. The other languages will be situated somewhere between them.

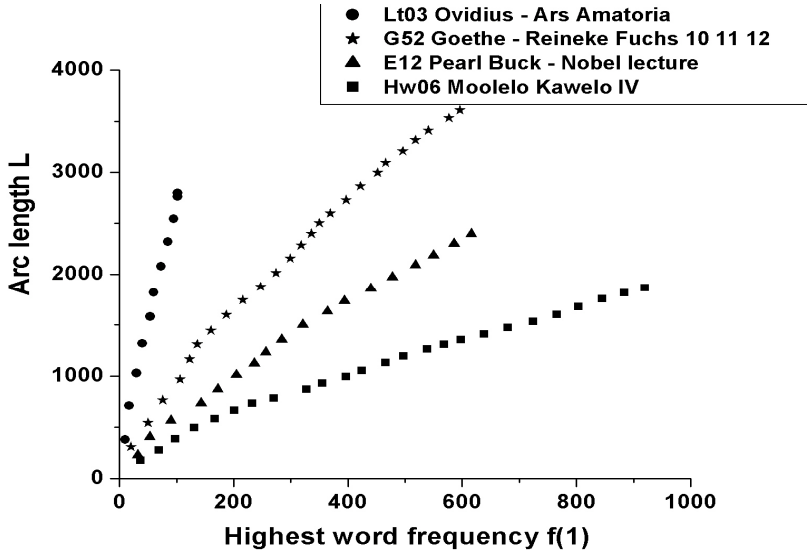


Figure 10. Arc length development in cumulative steps of 500 words in terms of f_1 for individual Hawaiian, English, German and Latin texts

In order to compare languages, one can determine confidence intervals for the curves taking many texts from every language. Figure 9 can be considered sufficiently representative for older German. Naturally one asks whether a certain aspect of the development of a

language can be studied comparing these curves for texts written in subsequent centuries. In any case the present approach opens a very broad view to possible future investigations.

In separate studies (Popescu, Altmann 2008a; Popescu, Altmann, Köhler 2008c) we have shown that the more synthetic a language the greater is the proportion of hapax legomena. Here we have shown that more synthetic languages have a longer arc than more analytic languages (cf. Figure 2). Consequently, the longer the arc, the more hapax legomena are in the text. As a matter of fact, this relationship is linear, as can be seen in Figure 11. Evidently, the reverse relationship holds, too. In Table 7 one can see the numerical values of this dependence.

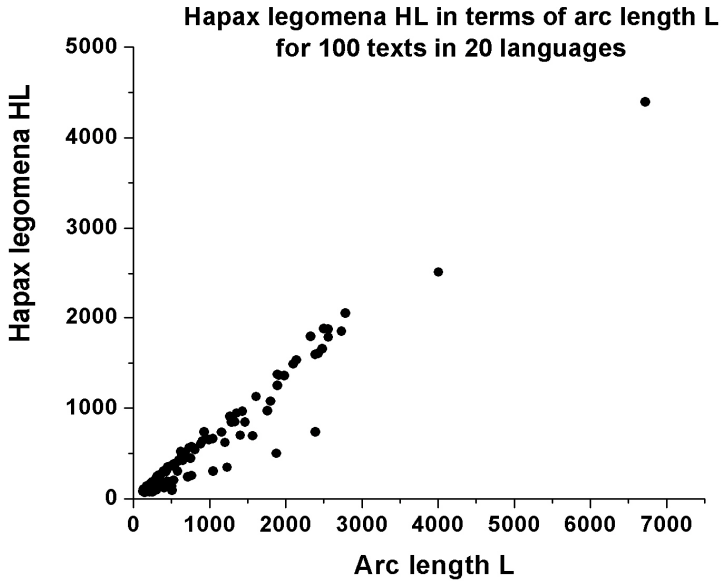


Figure 11. Dependence of hapax legomena on arc length

Table 7
Dependence of hapax legomena on arc length

ID	*L*	*HL*	B_5 = *HL/L*	*ID*	*L*	*HL*	B_5 = *HL/L*
B 01	428.45	298	0.696	Lk 03	317.63	174	0.548
B 02	205.38	153	0.745	Lk 04	125.56	80	0.637
B 03	289.80	212	0.732	Lt 01	2328.00	1792	0.770
B 04	297.03	222	0.747	Lt 02	2502.00	1878	0.751
B 05	247.30	187	0.756	Lt 03	2783.00	2049	0.736
Cz 01	684.17	517	0.756	Lt 04	1983.00	1359	0.685
Cz 02	586.22	412	0.703	Lt 05	930.00	737	0.792
Cz 03	1432.06	964	0.673	Lt 06	621.00	521	0.839
Cz 04	341.99	241	0.705	M 01	526.92	202	0.383
Cz 05	626.98	445	0.710	M 02	386.01	146	0.378
E 01	1042.85	662	0.635	M 03	384.62	133	0.346
E 02	1157.22	735	0.635	M 04	444.29	192	0.432
E 03	1204.91	620	0.515	M 05	715.18	239	0.334

E 04	1567.31	693	0.442	Mq 01	506.98	91	0.179
E 05	1760.86	971	0.551	Mq 02	178.59	86	0.482
E 07	1800.70	1075	0.597	Mq 03	500.37	138	0.276
E 13	2388.47	736	0.308	Mr 001	1612.43	1128	0.700
G 05	351.41	250	0.711	Mr 018	1890.34	1249	0.661
G 09	398.43	302	0.758	Mr 026	2098.93	1486	0.708
G 10	309.84	237	0.765	Mr 027	1467.65	846	0.576
G 11	306.80	232	0.756	Mr 288	2141.01	1534	0.716
G 12	175.44	141	0.804	R 01	886.35	606	0.684
G 14	132.54	107	0.807	R 02	1269.07	908	0.715
G 17	127.96	84	0.656	R 03	770.20	567	0.736
H 01	1288.83	844	0.655	R 04	764.36	573	0.750
H 02	907.18	638	0.703	R 05	599.19	424	0.708
H 03	332.44	259	0.779	R 06	451.75	353	0.781
H 04	674.06	509	0.755	Rt 01	315.91	127	0.402
H 05	314.40	250	0.795	Rt 02	264.75	128	0.483
Hw 03	764.27	255	0.334	Rt 03	255.86	98	0.383
Hw 04	1229.31	347	0.282	Rt 04	215.58	102	0.473
Hw 05	1047.48	302	0.288	Rt 05	250.69	73	0.291
Hw 06	1876.68	500	0.266	Ru 01	441.04	316	0.716
I 01	4007.01	2514	0.627	Ru 02	1356.70	946	0.697
I 02	2426.40	1604	0.661	Ru 03	1909.09	1365	0.715
I 03	534.33	382	0.715	Ru 04	2731.76	1850	0.677
I 04	1329.65	848	0.638	Ru 05	6722.04	4395	0.654
I 05	537.49	355	0.660	Sl 01	493.72	364	0.737
In 01	228.49	166	0.727	Sl 02	651.09	423	0.650
In 02	218.62	147	0.672	Sl 03	990.94	651	0.657
In 03	199.85	130	0.650	Sl 04	1404.13	701	0.499
In 04	217.37	145	0.667	Sl 05	2385.35	1593	0.668
In 05	195.65	121	0.618	Sm 01	403.17	119	0.295
Kn 003	1891.11	1373	0.726	Sm 02	303.92	96	0.316
Kn 004	733.26	564	0.769	Sm 03	168.39	75	0.445
Kn 005	2558.43	1784	0.697	Sm 04	214.17	76	0.355
Kn 006	2481.41	1655	0.667	Sm 05	149.49	66	0.442
Kn 011	2557.69	1873	0.732	T 01	680.99	465	0.683
Lk 01	184.77	127	0.687	T 02	807.46	540	0.669
Lk 02	579.97	302	0.521	T 03	748.50	447	0.597

In addition to the *B* indicators defined above, we can also introduce the hapax legomena proportion of the arc length by the simple indicator

(10) $$B_5 = \frac{Hapax\ Legomena\ HL}{Arc\ Length\ L}$$

theoretically positioned in the range <0,1>. The corresponding individual text values of B_5 are given in the last column of Table 7. Clearly, this indicator can also be used for typological purposes as shown in Figure 12 and Table 8 below.

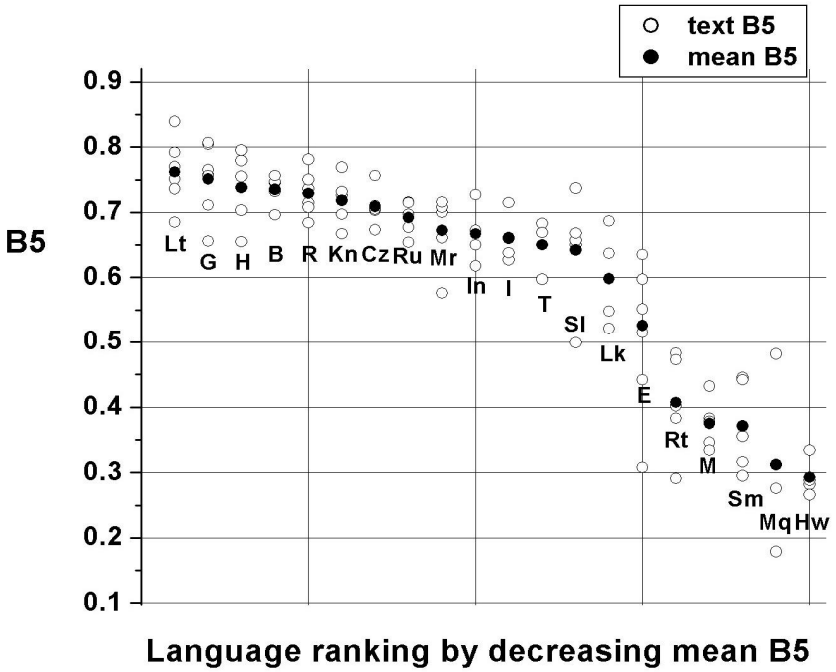


Figure 12. Indicator B_5 for 20 languages

Table 8
Means of B_5 for individual languages

1	Latin	0.762	11	Italian	0.660
2	German	0.751	12	Tagalog	0.650
3	Hungarian	0.738	13	Slovenian	0.642
4	Bulgarian	0.735	14	Lakota	0.598
5	Romanian	0.729	15	English	0.526
6	Kannada	0.718	16	Rarotongan	0.407
7	Czech	0.709	17	Maori	0.375
8	Russian	0.692	18	Samoan	0.371
9	Marathi	0.672	19	Marquesan	0.312
10	Indonesian	0.667	20	Hawaiian	0.293

Finally, an enlightening connection can be made between the relative arc length given by Eq. (3)

$$B_2 = \frac{L - L_{\min}}{L_{\max} - L_{\min}}$$

and the writer's view α radians (cf. Popescu, Altmann, 2007), defined as the angle seen from the *h*-point $P_3(h, h)$ and formed by the directions towards the distribution end $P_1(V, 1)$ and the top $P_2(1, f_1)$. In other words, this is the angle between the vectors **a** (a_x, a_y) and **b** (b_x, b_y) with the Cartesian components

$$a_x = -(h-1)$$
$$a_y = f_1 - h$$

and

$$b_x = V - h$$
$$b_y = -(h-1)$$

The expression of the cosine of this angle can easily be found as

$$\cos\alpha = -\frac{[(h-1)(f_1-h)+(h-1)(V-h)]}{[(h-1)^2+(f_1-h)^2]^{1/2}[(h-1)^2+(V-h)^2]^{1/2}} \tag{11}$$

slightly improving our preceding α values by about one percent. Table 9 and Figure 13 clearly demonstrate that the relative arc length B_2 merely represents a measure of the writer's view α.

Table 9
Dependence of B_2 on writer's view α radians

ID	V	f_1	h	α radians	B_2
B 01	400	40	10	1.885	0.763
B 02	201	13	8	2.558	0.470
B 03	285	15	9	2.527	0.430
B 04	286	21	8	2.090	0.618
B 05	238	19	7	2.060	0.588
Cz 01	638	58	9	1.745	0.835
Cz 02	543	56	11	1.808	0.809
Cz 03	1274	182	19	1.695	0.875
Cz 04	323	27	7	1.881	0.790
Cz 05	556	84	9	1.692	0.868
E 01	939	126	16	1.723	0.834
E 02	1017	168	22	1.735	0.837
E 03	1001	229	19	1.675	0.890
E 04	1232	366	23	1.653	0.911
E 05	1495	297	26	1.680	0.894
E 07	1597	237	25	1.699	0.861
E 13	1659	780	41	1.650	0.921
G 05	332	30	8	1.900	0.716
G 09	379	30	9	1.956	0.718
G 10	301	18	7	2.091	0.602
G 11	297	18	7	2.091	0.664
G 12	169	14	6	2.160	0.601

G 14	129	10	5	2.278	0.546
G 17	124	11	6	2.399	0.527
H 01	1079	225	12	1.633	0.939
H 02	789	130	8	1.637	0.925
H 03	291	48	4	1.649	0.915
H 04	609	76	7	1.668	0.885
H 05	290	32	6	1.778	0.837
Hw 03	521	277	26	1.721	0.851
Hw 04	744	535	38	1.697	0.871
Hw 05	680	416	38	1.726	0.847
Hw 06	1039	901	44	1.664	0.893
I 01	3667	388	37	1.683	0.877
I 02	2203	257	25	1.685	0.873
I 03	483	64	10	1.755	0.833
I 04	1237	118	21	1.791	0.798
I 05	512	42	12	1.944	0.648
In 01	221	16	6	2.058	0.590
In 02	209	18	7	2.100	0.647
In 03	194	14	6	2.156	0.553
In 04	213	11	5	2.178	0.583
In 05	188	16	8	2.328	0.599
Kn 003	1833	74	13	1.772	0.817
Kn 004	720	23	7	1.938	0.673
Kn 005	2477	101	16	1.752	0.829
Kn 006	2433	74	20	1.917	0.681
Kn 011	2516	63	17	1.912	0.696
Lk 01	174	20	8	2.141	0.632
Lk 02	479	124	17	1.754	0.812
Lk 03	272	62	12	1.830	0.749
Lk 04	116	18	6	2.011	0.630
Lt 01	2211	133	12	1.666	0.898
Lt 02	2334	190	18	1.677	0.895
Lt 03	2703	103	19	1.789	0.798
Lt 04	1910	99	20	1.817	0.757
Lt 05	909	33	8	1.852	0.704
Lt 06	609	19	7	2.044	0.760
M 01	398	152	18	1.742	0.836
M 02	277	127	15	1.749	0.846
M 03	277	128	17	1.775	0.823
M 04	326	137	15	1.730	0.854
M 05	514	234	26	1.742	0.836
Mq 01	289	247	22	1.742	0.831

Mq 02	150	42	10	1.909	0.698
Mq 03	301	218	14	1.680	0.893
Mr 001	1555	75	14	1.789	0.795
Mr 018	1788	126	20	1.759	0.827
Mr 026	2038	84	19	1.850	0.750
Mr 027	1400	92	21	1.860	0.755
Mr 288	2079	84	17	1.813	0.764
R 01	843	62	14	1.851	0.729
R 02	1179	110	16	1.742	0.836
R 03	719	65	12	1.791	0.820
R 04	729	49	10	1.810	0.766
R 05	567	46	11	1.867	0.744
R 06	432	30	10	2.015	0.731
Rt 01	223	111	14	1.766	0.819
Rt 02	214	69	13	1.842	0.730
Rt 03	207	66	13	1.855	0.738
Rt 04	181	49	11	1.887	0.719
Rt 05	197	74	15	1.881	0.706
Ru 01	422	31	8	1.883	0.679
Ru 02	1240	138	16	1.705	0.858
Ru 03	1792	144	21	1.743	0.825
Ru 04	2536	228	25	1.698	0.865
Ru 05	6073	701	41	1.638	0.926
Sl 01	457	47	9	1.796	0.829
Sl 02	603	66	13	1.814	0.753
Sl 03	907	102	13	1.718	0.840
Sl 04	1102	328	21	1.654	0.918
Sl 05	2223	193	25	1.724	0.849
Sm 01	267	159	17	1.747	0.825
Sm 02	222	103	15	1.796	0.770
Sm 03	140	45	13	2.024	0.624
Sm 04	153	78	12	1.814	0.760
Sm 05	124	39	11	2.002	0.664
T 01	611	89	14	1.764	0.802
T 02	720	107	15	1.742	0.830
T 03	645	128	19	1.763	0.811

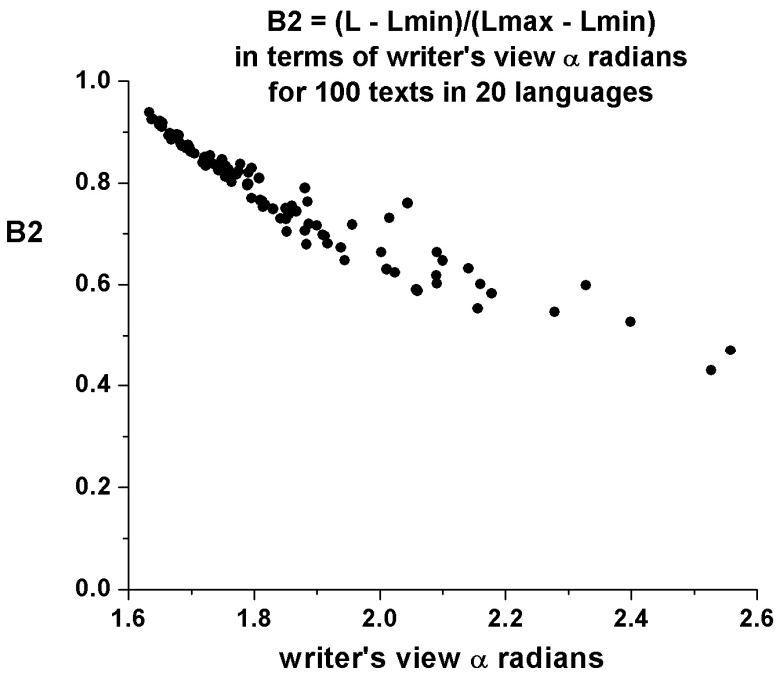


Figure 13. Indicator B_2 in terms of writer's view α for 100 texts in 20 languages

4. Arc length as a function of text indicators

In all the above dependence some empirical constants appeared whose linguistic interprêtation was not possible. As a matter of fact, capturing a dependence in this way is a play with the ceteris paribus conditions. The constants indicate that there is still something else having a constant influence on the dependent variable through the independent variable. In our case it can be shown that the arc length is a function of both V and f_1 as well as the h-point, all of which develop regularly with increasing N. Formally, we obtain

(12) $L \approx L_{max} - ch \approx V + f(1) - 1.3h$

telling us that the difference between the arc length L and its maximum value, L_{max} , approximated by the sum $(V + f(1))$, is always of the order of h. Though the multiplicative parameter c with h is rounded (to 1.3 from 1.297) and has a standard deviation of $s = 0.138$, this relationship seems to be a basic textual property involving the h-point.[3] In Table 10 the excellent agreement of observed (L) and computed (L_c) values can bee seen

[3] Actually, more exactly the relationship (12) should be written as

(13) $L = L_{max} - p(h - 1)$

where $L_{max} = (V - 1) + (f(1) - 1)$ is an ideal maximum arc length and p is also a constant of the order of unity (see more in the authors' coming book *New Aspects of Word Frequencies, Chapter 5. Arc length*, 2008).

Table 10
Observed and computed values of L

ID	*Text*	*N*	*V*	*f*(1)	*h*	*L*	L_c = *V*+*f*(1) - 1.3*h*
B 01	Boris 2 (Letter)	761	400	40	10	428.45	427.00
B 02	Ceneva1 (Letter)	352	201	13	8	205.38	203.60
B 03	Ceneva 2 (Letter)	515	285	15	9	289.8	288.30
B 04	Janko1 (Letter)	483	286	21	8	297.03	296.60
B 05	Janko 3 (Letter)	406	238	19	7	247.3	247.90
Cz 01	Hrabal 310, Expozé	1044	638	58	9	684.17	684.30
Cz 02	Hrabal 315, Lednová povídka:	984	543	56	11	586.22	584.70
Cz 03	Hrabal 316, Únorová povídka	2858	1274	182	19	1432.06	1431.30
Cz 04	Hrabal 319, Blitzkrieg	522	323	27	7	341.99	340.90
Cz 05	Hrabal 323, Protokol	999	556	84	9	626.98	628.30
E 01	Jimmy Carter, Nobel lecture	2330	939	126	16	1042.85	1044.20
E 02	Toni Morrison, Nobel lecture	2971	1017	168	22	1157.22	1156.40
E 03	George C. Marshall, Nobel lecture	3247	1001	229	19	1204.91	1205.30
E 04	J. M. Buchanan Jr., Nobel lecture	4622	1232	366	23	1567.31	1568.10
E 05	Saul Bellow, Nobel lecture	4760	1495	297	26	1760.86	1758.20
E 07	Sinclair Lewis, Nobel lecture	5004	1597	237	25	1800.7	1801.50
E 13	Richard Feynman, Nobel lecture	11265	1659	780	41	2388.47	2385.70
G 05	Goethe - Der Gott und die Bajadere	559	332	30	8	351.41	351.60
G 09	Goethe - Elegie 19	653	379	30	9	389.43	397.30
G 10	Goethe - Elegie 13	480	301	18	7	309.84	309.90
G 11	Goethe - Elegie 15	468	297	18	7	306.8	305.90
G 12	Goethe - Elegie 2	251	169	14	6	175.44	175.20
G 14	Goethe - Elegie 5	184	129	10	5	132.54	132.50
G 17	Goethe - Der Erlkönig	225	124	11	6	127.96	127.20
H 01	Orbán Viktor beszéde az Astoriánál	2044	1079	225	12	1288.83	1288.40
H 02	A nominalizmus forradalma	1288	789	130	8	907.18	908.60
H 03	Népszavazás	403	291	48	4	332.44	333.80
H 04	Egyre több	936	609	76	7	674.06	675.90
H 05	Kunczekolbász	413	290	32	6	314.4	314.20
Hw 03	Moolelo, Kawelo, Mokuna I	3507	521	277	26	764.27	764.20
Hw 04	Moolelo, Kawelo, Mokuna II	7892	744	535	38	1229.31	1229.60
Hw 05	Moolelo, Kawelo, Mokuna III	7620	680	416	38	1047.48	1046.60
Hw 06	Moolelo, Kawelo, Mokuna IV	12356	1039	901	44	1876.68	1882.80
I 01	Pellico, Le mie prigioni	11760	3667	388	37	4007.01	4006.90
I 02	Manzoni, I promessi sposi	6064	2203	257	25	2426.4	2427.50
I 03	Leopardi, Canti	854	483	64	10	534.33	534.00
I 04	Deledda, Canne al vento	3258	1237	118	21	1329.65	1327.70
I 05	de Amicis, Il cuore	1129	512	42	12	537.49	538.40
In 01	Assagaf-Ali Baba Jadi Asisten	376	221	16	6	228.49	229.20
In 02	BRI Siap Cetak Miliarder Dalam	373	209	18	7	218.62	217.90

In 03	Pengurus PSM Terbelah	347	194	14	6	199.85	200.20
In 04	Pemerintah Andalkan Hujan	343	213	11	5	217.37	217.50
In 05	Fajar	414	188	16	8	195.65	193.60
Kn 003	Pradhana Gurudhat	3188	1833	74	13	1891.11	1890.10
Kn 004	Pradhana Gurudhat	1050	720	23	7	733.26	733.90
Kn 005	T.R.Nagappa	4869	2477	101	16	2558.43	2557.20
Kn 006	T.R.Nagappa	5231	2433	74	20	2481.41	2481.00
Kn 011	D.N.S.Murthy	4541	2516	63	17	2557.69	2556.90
Lk 01	The fly on the window	345	174	20	8	184.77	183.60
Lk 02	Iktomi meets the prairie chicken	1633	479	124	17	579.97	580.90
Lk 03	Iktomi meets two women	809	272	62	12	317.63	318.40
Lk 04	Bean, grass, and fire	219	116	18	6	125.56	126.20
Lt 01	Vergil, Georgicon liber primus	3311	2211	133	12	2328	2328.40
Lt 02	Apuleius, Fables, Book 1	4010	2334	190	18	2502	2500.60
Lt 03	Ovidius, Ars amatoria, liber primus	4931	2703	103	19	2783	2781.30
Lt 04	Cicero, Post reditum in senatu oratio	4285	1910	99	20	1983	1983.00
Lt 05	Martialis, Epigrammata	1354	909	33	8	930	931.60
Lt 06	Horatius, Sermones.Liber 1 Sermo 1	829	609	19	7	621	618.90
M 01	Maori Nga Mahi	2062	398	152	18	526.92	526.60
M 02	Ko Te Paamu	1175	277	127	15	386.01	384.50
M 03	A Tawhaki	1434	277	128	17	384.62	382.90
M 04	Ka Pu Te Ruha	1289	326	137	15	444.29	443.50
M 05	Ka Kimi A Maui	3620	514	234	26	715.18	714.20
Mq 01	Story Kopuhoroto'e	2330	289	247	22	506.98	507.40
Mq 02	Ka´akai o Te Henua ´Enana	457	150	42	10	178.59	179.00
Mq 03	Te Hakamanu	1509	301	218	14	500.37	500.80
Mr 001	Prof. B.P.Joshi, Nisar Sheti	2998	1555	75	14	1612.43	1611.80
Mr 018	V.L.Pandy, Thumcha chehara	4062	1788	126	20	1890.34	1888.00
Mr 026	Kanchan Ganekar, Nath ha majha	4146	2038	84	19	2098.93	2097.30
Mr 027	Prof.Sarangar, Rashtriy Uthpann	4128	1400	92	21	1467.65	1464.70
Mr 288	Madhav Gadkari, Chaupher	4060	2079	84	17	2141.01	2140.90
R 01	Eminescu, Luceafarul	1738	843	62	14	886.35	886.80
R 02	Eminescu, Scrisoarea III	2279	1179	110	16	1269.07	1268.20
R 03	Eminescu, Scrisoarea IV	1264	719	65	12	770.2	768.40
R 04	Eminescu, Scrisoarea I	1284	729	49	10	764.36	765.00
R 05	Eminescu, Scrisoarea V	1032	567	46	11	599.19	598.70
R 06	Eminescu, Scrisoarea II	695	432	30	10	451.75	449.00
Rt 01	Kauraka Kauraka, Akamaramaanga	968	223	111	14	315.91	315.80
Rt 02	Tepania Puroku, Ko Paraka	845	214	69	13	264.75	266.10
Rt 03	Herekaiura Atama, Ko Tamaro	892	207	66	13	255.86	256.10
Rt 04	Temu Piniata, Te toa ko Teikapongi	625	181	49	11	215.58	215.70
Rt 05	Kaimaria Nikoro, Te toa ko Herehuaroa e Araitetonga	1059	197	74	15	250.69	251.50
Ru 01	Tolstoy, Metel’	753	422	31	8	441.04	442.60
Ru 02	Dostoevskij, Prestuplenie i nakaza-nie (p. I, ch. 1)	2595	1240	138	16	1356.7	1357.20

Ru 03	Pelevin, Buben verchnego mira	3853	1792	144	21	1909.09	1908.70
Ru 04	Turgenev, Bežin lug	6025	2536	228	25	2731.76	2731.50
Ru 05	Gogol, Portret	17205	6073	701	41	6722.04	6720.70
Sl 01	Jurčič, Sosedov sin	756	457	47	9	493.72	492.30
Sl 02	Levstik, Zveženj	1371	603	66	13	651.09	652.10
Sl 03	Grum, Vrata	1966	907	102	13	990.94	992.10
Sl 04	Cankar, V temi	3491	1102	328	21	1404.13	1402.70
Sl 05	Kočevar, Grof in menih	5588	2223	193	25	2385.35	2383.50
Sm 01	O le tala ia Sina ma lana tuna	1487	267	159	17	403.17	403.90
Sm 02	O le mea na maua ai le ava	1171	222	103	15	303.92	305.50
Sm 03	O upu faifai ma le gaoi	617	140	45	13	168.39	168.10
Sm 04	O le faalemigao	736	153	78	12	214.17	215.40
Sm 05	Ataliitaga	447	124	39	11	149.49	148.70
T 01	Hernandez, Magpinsan	1551	611	89	14	680.99	681.80
T 02	Hernandez, Limang Alas	1827	720	107	15	807.46	807.50
T 03	Rosales, Kristal Na Tubig	2054	645	128	19	748.5	748.30

Comparing the last two columns in Table 10 we see in Figure 14 the almost perfect coincidence yielding a determination coefficient of $R^2 = 1$. For further developments and verification of the relationship (12) see our coming book entitled "New Aspects of Word Frequencies", Chapter 5 Arc length (Popescu, Mačutek, Altmann 2008d).

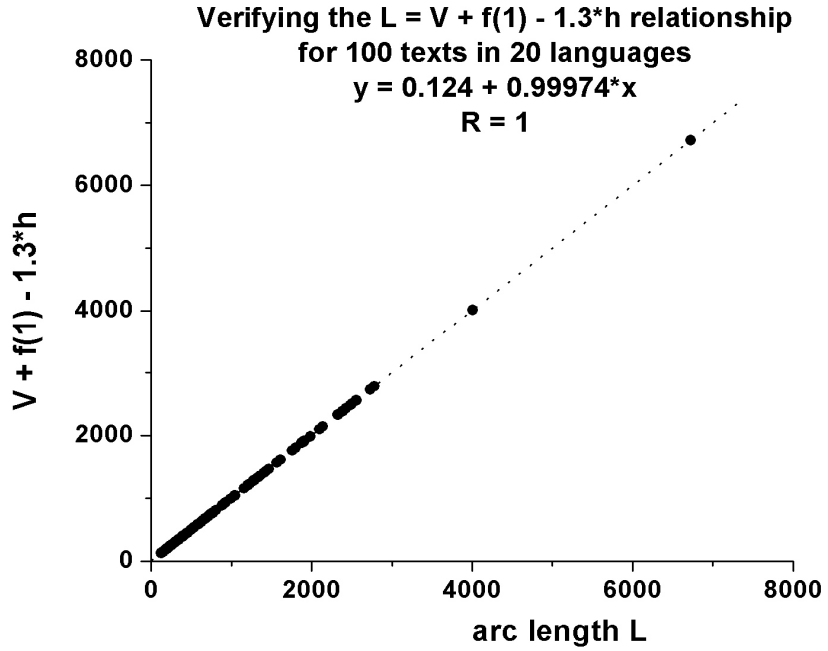


Figure 14. The coincidence of observed and computed arc length

5. Conclusion

Word frequencies deploy in texts very regularly. The deployment depends naturally on *N*, which influences *V*, the hapax legomena, f_1 and also *h* (cf. Popescu et al. 2008), but there are

still other boundary conditions which influence the arc. It is above all the morphological structure of language which prolongs or shortens *V* and the set of hapax legomena and in turn the arc length. On the other hand, arc length is also an indicator of redundancy. Since it depends also on f_1 representing the frequency of the most frequent word, it contributes strongly to the building of redundancy. As is well known, the greater f_1 the more the distribution deviates from uniformity, the smaller its entropy and the greater its redundancy. Further, since arc length depends on *V*, it is a kind of indicator of vocabulary richness And finally, it correlates strongly with the *writer´s view* of text development.

Thus, tracing the development and the final state of arc length can furnish us with a rich set of information on the text.

References

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

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 the human sciences: 514-546.* Berlin/New York: de Gruyter.

Krupa, V. (1965). On quantification of typology. *Linguistics 12, 31-36.*

Popescu, I.-I., Altmann, G. (2008a). Hapax legomena and language typology. *Journal of Quantitative Linguistics* (in print).

Popescu, I.-I., Altmann, G. (2008b). Zipf´s mean and language typology. *Glottometrics 16, 31-37.*

Popescu, I.-I., Altmann, G. Köhler, R. (2008c). Zipf´s law – another view. *Quality and Quantity* (submitted).

Popescu, I.-I., Mačutek J., Altmann, G. (2008d). *New Aspects of Word Frequencies, Chapter 5. Arc length* (pending)

Glottometrics 17, 2008, 43-50

Zur Diversifikation lateinischer und griechischer Hexameter

Karl-Heinz Best, Göttingen[1]

Abstract. The aim of this paper is to show that different forms of the Latin and Greek hexameter abide by the diversification law, too. The negative hypergeometric distribution can be fitted successfully.

Keywords: hexameter, diversification, Latin, Greek

Das Diversifikationsgesetz

Das von Altmann (1985) konzipierte Diversifikationsgesetz geht letztlich auf Zipfs Idee zurück, dass sich in der Sprache zwei Kräfte bemerkbar machen müssen: die *Force of Diversification*, die zusammen mit ihrem Widerpart, der *Force of Unification*, das Lexikon der Sprache gestaltet, indem sie danach trachtet, den Wortschatz und seine Bedeutungen zu vermehren, während die Unifikationskraft sie verringern will (Zipf 1949: 21). Dieses Prinzip des Ausgleichs zwischen den beiden Kräften gilt jedoch nicht nur für das Lexikon, sondern für alle Bereiche der Sprache. In Rothe (Hrsg. 1991) wurden eine ganze Reihe von solchen Anwendungsbereichen dargestellt und etliche auch mit Erfolg überprüft. Den Inhalt des Diversifikationsgesetzes kann man so umreißen: Wenn in einer Sprache irgendeine Entität semantisch oder formal in verschiedenen Versionen vorkommt, gibt es eine gesetzmäßige Beziehung zwischen den Häufigkeiten dieser Vorkommen. Die mathematische Ableitung dieses Gesetzes hat Altmann (1991) entwickelt; eine Überblicksdarstellung findet sich in Altmann (2005).

In diesem Beitrag soll es nun darum gehen, Daten zur Diversifikation lateinischer und griechischer Hexameter, die Drobisch (1866, 1868, 1872) in mehreren Untersuchungen vorstellt, daraufhin zu testen, ob sie dem Diversifikationsgesetz entsprechen. Eine einzelne Datei wurde bereits in Best (2008b) mit Erfolg getestet.

Zur Diversifikation von Hexametern

Die Verteilung der Hexameterformen hat schon einmal Grotjahn (1979: 205-210; passim) einer Analyse unterzogen. Dabei bildete er allerdings nur drei Gruppen: Vergil, vergilianische und nichtvergilianische Dichter. Es geht ihm um den Trend, dass Daktylen eine Tendenz zu den vorderen Versfüßen zeigen, die einer Regressionsgleichung folgt.

Im Folgenden werden die Daten Drobischs anders behandelt: Das Vorkommen der Hexameterformen wird als ein Diversifikationsphänomen behandelt; die ausgewählten Texte werden je für sich daraufhin getestet, ob sie einem der Modelle für das Diversifikationsgesetz folgen.

Hauptsächlich in zwei Untersuchungen hat sich Drobisch (1866, 1868) mit den lateinischen und griechischen Hexametern befasst. Dabei behandelte er in Drobisch (1866) 16 Texte von 15 lateinischen Autoren und berücksichtigte dabei relativ kurze Textabschnitte, meist 560 Verse je Autor. Einige dieser Autoren greift er in Drobisch (1868) wieder auf, dies-

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

mal mit wesentlich längeren Textabschnitten. Die folgenden Tabellen enthalten alle Autoren, die in einer dieser beiden Abhandlungen vorkommen; diejenigen, die in Drobisch (1868) erneut behandelt werden, jedoch mit längeren Textausschnitten, werden auch aus dieser zweiten Arbeit genommen. Hinzu kommen die drei griechischen Texte, die nur in Drobisch (1868) vorgestellt werden, sowie ein weiterer aus Drobisch (1872). Zunächst die lateinischen Hexameter, mit denen sich der Autor am meisten befasst hat. Gegenstand der Untersuchungen waren immer die ersten vier Füße jedes Verses.

Die lateinischen Hexameter

Nach den Erfahrungen mit den deutschen Hexametern stellt sich die Frage danach, welches Modell für die der Antike geeignet sein sollte. Bei den deutschen Hexametern wurden in Anlehnung an die Untersuchung von Schweers & Zhu (1991) zu Wortartenverteilungen die Pólya-Verteilung und die negative hypergeometrische Verteilung als Modelle verwendet (Best 2008a). Wie sich gezeigt hat, kann die 1-verschobene negative hypergeometrische Verteilung

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

als Modell für alle lateinischen und griechischen Texte verwendet werden, wie die folgenden Tabellen zeigen. In den Tabellen werden folgende Symbole benutzt:

d = Daktylus; *s* = Spondeus
x = Rang der Hexameterformklasse
n_x = beobachtete Häufigkeit der Hexameterform
NP_x = aufgrund der 1-verschobenen negativen hypergeometrischen Verteilung berechnete Häufigkeit der Hexameterform
K, M, n = Parameter der negativen hypergeometrischen Verteilung
FG = Freiheitsgrade
X^2 = Chiquadrat
P = Überschreitungswahrscheinlichkeit des Chiquadrats
$C = X^2/N$ = Diskrepanzkoeffizient

An den Werten für die Testkriterien *P* und *C* kann man ablesen, ob die gewählte Verteilung für die Daten eines bestimmten Textes ein geeignetes Modell ist. Die Anpassung wird als erfolgreich angesehen, wenn $P \geq 0.05$ oder $C \leq 0.01$; als noch akzeptabel werden Ergebnisse mit $0.02 \leq C < 0.01$ gewertet. Bei Dateien mit kleinem Umfang ist *P* das geeignetere Kriterium; *C* wird entsprechend bei umfangreichen Dateien verwendet oder wenn $FG = 0$.

Tabelle 1
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1868: 26, 20, 29)

	Vergil, *Aeneis*			Vergil, *Georgica*			Vergil, *Aeneis* u. *Georgica*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsss	423	458.95	dsss	344	374.54	dsss	767	828.68

2	ddss	338	339.07	dsds	262	258.11	dsds	587	595.70
3	dsds	325	288.03	ddss	249	213.18	ddss	587	500.60
4	sdss	297	254.73	sdss	213	185.40	sdss	510	440.01
5	ddds	229	229.27	dssd	152	164.99	ddds	365	394.48
6	ssss	191	208.08	ddds	136	148.53	dssd	319	357.12
7	ssds	170	189.49	ssss	128	134.47	ssss	319	324.72
8	dssd	167	172.55	ssds	119	121.95	ssds	289	295.48
9	sdds	167	156.68	sdds	111	110.44	sdds	278	268.29
10	ddsd	136	141.43	ddsd	105	99.56	ddsd	241	242.35
11	dsdd	117	126.47	sdsd	88	89.03	dsdd	195	217.02
12	sdsd	106	111.45	dsdd	70	78.58	sdsd	194	191.69
13	sssd	102	96.01	sssd	56	67.93	sssd	158	165.69
14	dddd	66	79.60	ssdd	52	56.66	dddd	115	138.08
15	sddd	60	61.32	dddd	49	44.10	ssdd	110	107.20
16	ssdd	58	38.87	sddd	42	28.52	sddd	102	68.91
Σ		2952			2176			5136	
	K = 2.3190 M = 0.7660 n = 15 C = 0.0109			K = 2.2298 M = 0.7129 n = 15 C = 0.0119			K = 2.2733 M = 0.7442 n = 15 C = 0.0117		

Tabelle 2
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1868: 32; 1866: 93, 95)

	Vergil, *Bucolica*			Ennius, Fragmente			Cicero, Arat-Übersetzung		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	ddss	107	115.71	ssss	64	61.48	dsss	92	97.12
2	dsss	90	86.04	sdss	39	42.88	sdss	77	74.36
3	dsds	79	73.84	dsss	39	36.06	ssss	74	62.80
4	dssd	64	66.09	ssds	35	32.02	ddss	74	54.46
5	sdss	63	60.28	sssd	25	29.16	sdds	40	47.67
6	ddsd	59	55.53	dssd	24	26.94	ssds	35	41.79
7	ddds	57	51.42	dsds	24	25.09	dsds	34	36.52
8	dsdd	43	47.69	ddss	24	23.48	ddds	33	31.72
9	sdds	43	44.22	sdds	23	22.04	sdsd	20	27.26
10	ssss	40	40.87	ddds	21	20.70	sssd	18	23.09
11	sdsd	38	37.56	ddsd	20	19.40	dssd	16	19.16
12	ssds	29	34.20	dddd	20	18.12	ddsd	16	15.45
13	dddd	27	30.67	sdsd	19	16.78	ssdd	10	11.93
14	sssd	26	26.78	sddd	15	15.32	sddd	9	8.61
15	ssdd	23	22.19	dsdd	12	13.56	dsdd	8	5.48
16	sddd	21	15.91	ssdd	10	10.96	dddd	4	2.58
Σ		809			414			560	
	K = 2.1347 M = 0.7621 n = 15 FG = 12 X^2 = 5.64 P = 0.93			K = 1.9199 M = 0.7074 n = 15 FG = 12 X^2 = 2.77 P = 0.99			K = 2.9229 M = 0.8219 n = 15 FG = 12 X^2 = 17.79 P = 0.12		

Tabelle 3
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1866: 97, 99; 1868: 33)

	Lukrez, *de rer. nat.*			Catull, 2 Gedichte			Horaz, *Satiren*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsss	88	95.01	dsss	124	114.78	dsss	285	301.27
2	ddss	72	71.17	sdss	65	71.51	sdss	228	228.65
3	sdss	63	60.00	ddss	55	53.64	dsds	205	197.34
4	ddds	56	52.29	dsds	51	42.23	ddss	203	176.85
5	dsds	51	46.17	ssss	43	33.84	ssss	174	161.14
6	ssss	39	40.95	ssds	25	27.26	ssds	137	148.03
7	sdds	37	36.31	dssd	15	21.90	dssd	134	136.48
8	dssd	36	32.07	ddds	15	17.45	ddds	115	125.91
9	ssds	26	28.11	sdsd	8	13.73	sdds	108	115.93
10	sssd	21	24.37	sdds	7	10.61	sssd	102	106.27
11	sdsd	17	20.79	ddsd	6	8.00	sdsd	93	96.67
12	ddsd	17	17.31	dsdd	5	5.83	ddsd	92	86.91
13	dsdd	12	13.92	sddd	4	4.05	dsdd	79	76.67
14	sddd	10	10.57	sssd	3	2.63	ssdd	63	65.52
15	dddd	8	7.20	ssdd	3	1.54	dddd	48	52.62
16	ssdd	7	3.76	dddd	1	1.00	sddd	46	35.73
Σ		560			430			2112	
	K = 2.6831 M = 0.7934 n = 15 FG = 12 X^2 = 6.53 P = 0.89			K = 3.7124 M = 0.7013 n = 16 FG = 11 X^2 = 13.49 P = 0.26			K = 2.2333 M = 0.7818 n = 15 FG = 12 X^2 = 12.57 P = 0.40		

Tabelle 4
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch (1868: 36, 38; 1866: 104)

	Horaz, *Episteln*			Horaz, *Satiren* u. *Episteln*			Ovid, *Metamorphosen*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsss	237	239.42	dsss	522	538.23	dssd	78	73.70
2	sdss	198	197.94	sdss	426	424.88	dsss	76	72.85
3	dsds	189	177.57	dsds	394	373.41	dsds	63	67.63
4	ddss	168	163.17	ddss	371	338.74	ddss	60	61.10
5	dssd	147	151.48	ssss	299	311.57	ddsd	57	54.16
6	ddsd	132	141.25	dssd	281	288.49	dsdd	54	47.22
7	ssss	125	131.87	ssds	261	267.84	ddds	45	40.51
8	ssds	124	122.97	sdsd	225	248.66	dddd	33	34.15
9	ddds	109	114.32	ddds	224	230.33	sdss	26	28.25
10	sdds	109	105.69	sdds	217	212.36	sdsd	21	22.87
11	ddsd	97	96.92	sssd	191	194.32	sdds	13	18.03
12	dsdd	94	87.78	ddsd	189	175.75	sddd	11	13.76
13	sssd	89	78.00	dsdd	173	156.07	ssss	6	10.08

14	ssdd	59	67.13	ssdd	122	134.38	ssds	6	7.00
15	sddd	54	54.31	sddd	100	108.94	ssdd	6	4.50
16	dddd	36	37.19	dddd	84	75.02	sssd	5	4.18
Σ		1967			4079			560	
	K = 2.2973 M = 0.8513 n = 15 FG = 12 X^2 = 5.37 P = 0.94			K = 2.2462 M = 0.8122 n = 15 FG = 12 X^2 = 13.94 P = 0.30			K = 4.3902 M = 1.1204 n = 17 FG = 12 X^2 = 7.13 P = 0.85		

Tabelle 5
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1866: 107, 110, 112)

	Manilius, *Astronomica*			Persius, *Satiren*			Juvenal, *Satiren*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsss	93	98.41	dsss	118	123.49	dsss	85	93.84
2	sdss	67	68.29	ddss	96	84.52	sdss	67	64.88
3	dsds	60	56.35	dsds	68	68.82	dsds	64	53.80
4	ddss	57	48.84	sdss	62	58.82	ddss	51	46.98
5	ssss	48	43.26	dssd	48	51.31	dssd	40	42.00
6	ssds	34	38.71	ddds	39	45.19	ssds	38	38.00
7	dssd	33	34.80	sdsd	35	39.91	ddds	35	34.59
8	ddsd	30	31.31	ddsd	35	35.21	ddsd	29	31.56
9	sdds	28	28.08	dsdd	32	30.92	sdsd	26	28.77
10	sdsd	22	25.04	ssss	30	26.91	ssss	25	26.14
11	ddds	22	22.11	ssds	27	23.10	sdds	22	23.58
12	dsdd	19	19.21	sdds	19	19.43	sddd	20	21.03
13	ssdd	16	16.29	dddd	14	15.83	dsdd	19	18.41
14	sssd	11	13.26	sssd	12	12.25	ssdd	14	15.62
15	dddd	11	9.96	sddd	9	8.59	sssd	13	12.44
16	sddd	9	6.08	ssdd	5	4.70	dddd	12	8.36
Σ		560			649			560	
	K = 2.3300 M = 0.7220 n = 15 FG = 12 X^2 = 5.43 P = 0.94			K = 2.5145 M = 0.7207 n = 15 FG = 12 X^2 = 4.97 P = 0.96			K = 2.1720 M = 0.7126 n = 15 FG = 12 X^2 = 5.76 P = 0.93		

Tabelle 6
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1866: 114, 116, 118)

	Lukan, *Pharsalia*			Silius Italicus, *Punica*			Valerius Flaccus, *Argonautica*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsds	98	109.62	dsss	75	76.89	dsds	131	123.17
2	dsss	83	71.36	sdss	63	62.42	ddss	75	86.52
3	ddss	59	57.29	ssds	54	55.06	ddds	64	70.69
4	sdss	58	48.76	ssss	53	49.72	dsss	63	60.22

5	dssd	39	42.59	dsds	47	45.33	ddsd	54	52.17
6	ddds	33	37.67	ddss	47	41.45	dssd	52	45.49
7	ssds	32	33.53	sssd	34	37.88	dsdd	49	39.69
8	ddsd	29	29.88	sdsd	32	34.52	ssds	30	34.51
9	sdds	28	26.56	dssd	28	31.27	sdsd	24	29.77
10	sdsd	23	23.48	ddds	26	28.08	dddd	24	25.39
11	dsdd	19	20.55	sdds	25	24.90	sdss	22	21.27
12	ssss	15	17.70	ddsd	24	21.68	sdds	21	17.38
13	dddd	13	14.86	ssdd	18	18.36	ssss	12	13.67
14	sssd	12	11.96	dsdd	16	14.86	sddd	11	10.10
15	sddd	11	8.87	dddd	11	11.04	ssdd	5	6.66
16	ssdd	8	5.31	sddd	7	6.57	sssd	3	3.31
Σ		560			560			640	
	$K = 2.3141$ $M = 0.6785$ $n = 15$ $FG = 12$ $X^2 = 8.64$ $P = 0.73$			$K = 2.5108$ $M = 0.8477$ $n = 15$ $FG = 12$ $X^2 = 2.54$ $P = 0.99$			$K = 2.7286$ $M = 0.7484$ $n = 15$ $FG = 12$ $X^2 = 9.26$ $P = 0.68$		

Tabelle 7
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1866: 120, 123; 1868: 41)

	Statius, *Thebais*			Claudian, *raptus Proserpinae*			Leibniz, *Epicedium*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dsds	83	91.14	dsds	102	99.64	dsss	59	65.63
2	dsss	76	67.82	ddss	83	85.89	ddss	57	45.18
3	ddds	57	57.40	dsss	75	73.54	dsds	33	36.71
4	ddss	53	50.40	sdss	67	62.41	ddds	31	31.25
5	ssds	43	44.93	ssds	51	52.42	sdss	30	27.11
6	dsdd	40	40.32	ddds	38	43.50	dssd	24	23.71
7	ddsd	39	36.23	sdds	34	35.60	sdds	20	20.77
8	sdss	34	32.49	dssd	30	28.66	dsdd	16	18.16
9	sdds	32	28.99	sdsd	24	22.64	ssss	13	15.78
10	dssd	24	25.65	ddsd	21	17.47	ddsd	12	13.56
11	dddd	17	22.40	dsdd	14	13.12	sdsd	10	11.48
12	sdsd	16	19.19	ssdd	8	9.51	sddd	10	9.49
13	ssss	14	15.97	dddd	5	6.59	ssds	10	7.58
14	ssdd	14	12.66	sssd	3	4.30	ssdd	6	5.71
15	sssd	10	9.16	sddd	3	2.58	dddd	5	3.87
16	sddd	8	5.25	ssss	2	2.12	sssd	2	2.00
Σ		560			640			338	
	$K = 2.4992$ $M = 0.7798$ $n = 15$ $FG = 12$ $X^2 = 6.37$ $P = 0.90$			$K = 4.6521$ $M = 0.9967$ $n = 17$ $FG = 12$ $X^2 = 3.33$ $P = 0.99$			$K = 2.6269$ $M = 0.7295$ $n = 15$ $FG = 12$ $X^2 = 6.75$ $P = 0.87$		

Die griechischen Hexameter

Tabelle 8
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1868: 44, 50, 57)

	Homer, *Ilias*			Homer, *Odyssee*			Theokrit, *1. Idylle*		
x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x	Hexame-terform	n_x	NP_x
1	dddd	350	365.83	dddd	410	423.70	dsdd	31	31.55
2	dsdd	320	294.25	dsdd	323	309.41	dddd	21	18.75
3	sddd	296	247.88	sddd	277	252.61	sddd	13	14.35
4	ddds	196	210.74	ddds	185	212.35	dssd	11	11.75
5	sdds	155	179.02	ssdd	176	180.20	ssdd	10	9.89
6	ssdd	149	151.22	sdds	161	153.05	ddsd	9	8.41
7	dsds	145	126.58	dsds	149	129.39	ssds	7	7.17
8	ssds	78	104.65	ddsd	92	108.42	dsds	5	6.07
9	ddsd	76	85.14	ssds	82	89.64	sdds	5	5.08
10	sdsd	73	67.88	dssd	71	72.75	sssd	5	4.14
11	dssd	56	52.71	sdsd	65	57.54	sdsd	4	3.23
12	sdss	28	39.55	sssd	35	43.91	ddds	2	2.30
13	ddss	22	28.32	ddss	34	31.78	dsss	1	1.31
14	dsss	21	18.97	sdss	20	21.17			
15	sssd	19	11.48	dsss	18	12.16			
16	ssss	8	7.78	ssss	5	4.93			
Σ		1992			2103			124	
	K = 3.8259 M = 0.9011 n = 16 C = 0.0188			K = 3.2340 M = 0.8001 n = 15 FG = 12 X^2 = 19.39 P = 0.08			K = 2.3346 M = 0.6292 n = 12 FG = 9 X^2 = 1.17 P = 0.99		

Tabelle 9
Anpassung der 1-verschobenen negativen hypergeometrischen Verteilung
(Quelle: Drobisch 1872: 10)[2]

Theognis, *Elegische Dichtungen*							
x	Hexameterform	n_x	NP_x	*x*	Hexameterform	n_x	NP_x
1	dsdd	117	123.61	9	ssds	26	26.26
2	sddd	99	87.77	10	ddsd	25	21.62
3	dddd	78	71.18	11	sdsd	23	17.39
4	ssdd	66	59.83	12	sssd	21	13.54
5	dsds	38	50.94	13	ddss	7	10.05
6	dssd	36	43.51	14	sdss	4	6.91
7	sdds	34	37.08	15	dsss	3	4.14
8	ddds	28	31.38	16	ssss	2	1.78
Σ	607						
	K = 3.0579 M = 0.7710 n = 15 FG = 12 X^2 = 17.22 P = 0.14						

[2] Hultgren, dessen Ergebnisse Drobisch in seinem Beitrag vorstellt, verweist darauf, dass die hier als 1. Buch des Theognis aufgeführten Texte nicht alle von einem und demselben Autor stammen können (Drobisch 1872: 9).

Ergebnis

Wie sich gezeigt hat, kann in allen Fällen die 1-verschobene negative hypergeometrische Verteilung als Modell angewendet werden, bei kürzeren Textabschnitten immer mit guten Ergebnissen, bei den längeren mit etwas weniger guten. Der Grund dafür liegt vermutlich in den bekannten Problemen der Inhomogenität von Daten, die bei willkürlicher Bildung von Textabschnitten, bei Mischungen von Stichproben und bei größeren Ausschnitten auftreten können (Altmann 1992). Umso höher ist das Gesamtergebnis zu bewerten, wenn trotzdem ein gemeinsames Modell erfolgreich erprobt werden konnte. Das Diversifikationsgesetz hat sich damit in einer weiteren Domäne als erfolgreich erwiesen.

Literatur

Altmann, Gabriel (1985). Semantische Diversifikation. *Folia Linguistica XIX, 177-200.*

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

Altmann, Gabriel (1992). Das Problem der Datenhomogenität. In: Rieger, Burghard (Hrsg.), *Glottometrika 13* (S. 287-298). Bochum: Brockmeyer.

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 646-658). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2008a). Zur Diversifikation deutscher Hexameter. Mskr. (Eingereicht)

Best, Karl-Heinz (2008b). Moritz Wilhelm Drobisch (1802-1896). (In diesem Heft)

Drobisch, Moritz Wilhelm (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 18, 75-139.*

Drobisch, Moritz Wilhelm (1868). Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 20, 138-160.*

Drobisch, Moritz Wilhelm (1872). Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 24, 1-33.*

Grotjahn, Rüdiger (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum: Brockmeyer.

Schweers, Anja, & Zhu, Jinyang (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, Ursula (Hrsg.), *Diversification processes in language: grammar* (S. 157-165). Hagen: Rottmann..

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

Glottometrics 17, 2008, 51-65

A mathematical analysis of parts-of-speech systems

Relja Vulanović
North Canton, Ohio[1]

Abstract. The classification of parts-of-speech systems, introduced in Hengeveld (1992), is generalized to include other theoretical possibilities, regardless of whether they are realized in natural languages or not. In the new classification, there are three flexible parts-of-speech types which contain several possible subtypes. Those are analyzed from the point of view of grammar efficiency. It is shown that the most efficient systems have free word order and use the same word class for both the head of the predicate phrase and the head of the referential phrase. This structure is not attested in natural languages. As for attested systems, they occupy the highest position on the grammar-efficiency scale only in some special cases and when word order is fixed.

Keywords: Word classes, propositional functions, parts-of-speech systems, grammar efficiency

1. Introduction

This paper is concerned with the classification of parts-of-speech (PoS) systems introduced in Hengeveld (1992) and further discussed in Hengeveld, Rijkhoff & Siewierska (2005) – HRS from this point on, Hengeveld & Rijkhoff (2005), and Rijkhoff (2007). A functional approach is taken in these four papers: parts of speech are defined based on their propositional functions. Four propositional functions are considered: P – the obligatory head of the predicate phrase (PPh), R – the obligatory head of the referential (noun) phrase (RPh), p – the optional modifier of PPh, and r – the optional modifier of RPh. A verb (V) is a word that can only be used as the head of a PPh, a noun (N) is a word that can be used as the head of a RPh, an adjective (a) is a word that can be used as a modifier of the RPh head, and a manner adverb (m) is a word that can be used as a modifier of the PPh head (the approach is not extended to other types of adverbs). The above-mentioned uses are the typical ones which define a PoS system. However, not all languages have PoS systems in which there are four word classes to indicate the four propositional functions. A sample of 50 languages in HRS shows that some languages have fewer than four propositional functions and also that some languages have *flexible* PoS systems, in which the number of word classes is less than the number of propositional functions. Flexible languages use other word classes like modifier (M), non-verb (Λ), or contentive (C). What these classes mean, can be easily concluded from Table 1 which shows the HRS classification of PoS system types. Since another classification is going to be proposed here, the HRS PoS system types are labeled HRS1-HRS7.

[1] Address correspondence to: Relja Vulanović, Department of Mathematical Sciences, Kent State University Stark Campus, 6000 Frank Ave NW, North Canton, OH 44720, USA.
E-mail: rvulanov@kent.edu

Table 1
The HRS typology of PoS systems based on
what word classes have what propositional functions

type	P	R	r	p
HRS1	C			
HRS2	V	Λ		
HRS3	V	N	M	
HRS4	V	N	a	m
HRS5	V	N	a	-
HRS6	V	N	-	-
HRS7	V	-	-	-

PoS systems types HRS1, HRS2, and HRS3 are flexible and the remaining four types are *rigid.* In a rigid system, the number of word classes is the same as the number of propositional functions.

The following examples from Turkish, due to Göksel & Kerslake (2005: 49) and taken here from Hengeveld & van Lier (to appear), show the situation in a PoS system of type HRS2:

güzel-im
beauty-1POSS
'my beauty'

güzel bir kopek
beauty ART dog
'a beautiful dog'

güzel konuştu
beauty s/he.spoke
's/he spoke well'

This is meant to illustrate that Turkish has a whole class of words like 'güzel' which function as R, r, or p. It is customary to call this class *non-verbs* and I conveniently denote it by Λ. Therefore, in an HRS2 system, we have the following assignment of word classes to propositional functions:

$V \rightarrow P, \quad \Lambda \rightarrow R, r, p.$

Examples illustrating and motivating other PoS system types can be found in HRS and the above-mentioned related papers.

The HRS classification of PoS types contains also languages of intermediate types. In fact, Turkish is classified in HRS as a language with PoS type between HRS2 and HRS3. Another example is Mundari (between types HRS1 and HRS2), which is discussed in detail in Hengeveld and Rijkhoff (2005). Intermediate PoS systems are not considered in the present paper.

The HRS classification of PoS systems has been subjected to some criticism; see Croft (2000) and Baker (2003) for instance. This is not surprising in view of the fact that there is no unified linguistic definition of word classes; a survey of different approaches can be found in Baker (2003) and Rijkhoff (2007), in addition to HRS. In his book, Baker defines and analyzes verbs, nouns, and adjectives within the framework of generative grammar. His results are different from those in the work of Hengeveld and his collaborators. For instance, Baker does not consider adverbs separate from adjectives and finds that verbs, nouns, and adjectives are categories universally shared by natural languages. Therefore, according to Baker, some of the parts of speech in Table 1 and some PoS types are a myth. As a mathematician, I will leave these issues to linguists to resolve. The HRS classification still seems to be a viable one and I would like to discuss it further in the present paper.

One striking feature of Table 1 is the absence of flexible systems with less than four propositional functions. This is like stating that a PoS system has to be rigid if has one, two, or three propositional functions. However, even the HRS sample contains a language, Tagalog, which is described there as a type HRS1 language without p. This motivates a classification based on dual numbering: a type *k.n* PoS system would mean a system with *k* word classes and *n* propositional functions (n = 1, 2, 3, or 4 and k = 1,..., n). Therefore, Tagalog would have a 1.3 PoS system, as opposed to Samoan, the other type HRS1 language in the HRS sample, which is 1.4 in the notation I propose here. Table 2 shows the place of the HRS types in the new classification. No PoS system type is possible below the diagonal of Table 2 since there $k > n$. Rigid systems are on the diagonal, where $k = n$, and flexible systems are above the diagonal. There are three flexible types (1.2, 1.3, and 2.3) which are not covered by the HRS classification. They represent PoS systems that are at least theoretically possible.

Table 2
The position of HRS PoS system types within the dual-numbering classification
(+ indicates a type missing in the HRS classification;
* the HRS type is just a subtype of the new classification)

$k \setminus n$	1	2	3	4
1	HRS7	+	+	HRS1
2	-	HRS6	+	HRS2*
3	-	-	HRS5*	HRS3*
4	-	-	-	HRS4

I am indeed interested here in all theoretical possibilities regarding how *n* propositional functions can be fulfilled by *k* parts of speech. This includes other ways of assigning word classes to the propositional functions, not just those presented in Table 1. For instance, the assignment

$$V \rightarrow P, \quad \Lambda \rightarrow R, r, p$$

in type 2.4, which has already been mentioned above, is only one possibility and, therefore just one subtype of type 2.4. One of the other possible assignments would be

$$H \rightarrow P, R, \quad M \rightarrow r, p, \qquad (1)$$

where H denotes a word class that can typically function as either head. This means that my 2.4 class is not equivalent to HRS2, the latter being just one subtype of 2.4. Also, types HRS3 and HRS5 are subtypes of 3.4 and 3.3 respectively. All possible subtypes are presented in section 2, regardless of whether there is a natural language with such a PoS system or not. The possibilities I explore are not in Newmeyer's (2005) sense of languages that are permissible by Universal Grammar – they are simply all mathematical possibilities. Linguists, although usually not interested in *all* possibilities, do discuss structures that cannot be found in natural languages. Baker (2003), for instance, devotes enough pages to languages that he otherwise considers hypothetical, viz. languages without verbs, or nouns, or adjectives. Even the HRS classification is hypothetical to some extent, to quote Rijkhoff (2007, p. 718), "[b]ecause languages are dynamic entities, they can only approximate the ideal types in this classification." Also, type HRS7 is not attested in natural languages but some Native American languages, like Tuscarora, come close to it. Tuscarora is classified in HRS as an intermediate type language between HRS6 and HRS7.

Some PoS types, like 1.4 for instance, have no subtypes. In the case of type 1.4 (HRS1), there is only one way of assigning C to the four propositional functions: C → P, R, r, p. On the other hand, three flexible types, 2.3, 2.4, and 3.4, have essentially different subtypes. The question, then, is how the subtypes within one type compare to each other with respect to their efficiency. This is in the sense of grammar efficiency of Vulanović (1991, 1993, 2003, 2007), which is briefly reviewed here in section 3. It is shown that the complete efficiency formula is not necessary and that the *parsing ratio* (cf. Vulanović, to appear b) satisfies the needs of this paper. After the parsing ratio is discussed in section 3, efficiency results for the subtypes are presented in section 4. One of the conclusions is that the attested PoS systems have the highest efficiency only in some special cases. In each of the three types with subtypes, the most efficient system is the same kind of unattested structure: word order is completely free and one word class functions as both heads (like in (1) above).

This paper is a continuation of my two recent works on PoS systems. Both Vulanović (to appear a) and (to appear b) discuss types HRS1-HRS4. In Vulanović (to appear a), I am interested in the functional ambiguity that flexible languages are prone to. I use a formal combinatorial analysis to show that PoS systems of type HRS1 and HRS2 cannot resolve this ambiguity without syntactic or morphological markers. However, a further analysis of the flexible languages in the HRS sample reveals that many of them tolerate some amount of ambiguity. They either have no markers at all, or they use them with some degree of redundancy, and, therefore, ambiguous structures are not eliminated. The other paper deals with most of the flexible languages in the HRS sample and some additional natural (mainly type HRS4) and hypothetical languages. Each linguistic structure in the sample is represented formally and three types of grammar efficiency are calculated. The main conclusion is that grammar efficiency of natural languages is much smaller than what is theoretically possible. This is so because, on the one hand, word order is not as free as possible, and, on the other hand, word classes, together with markers, are not assigned to the propositional functions in the most efficient way.

Thus, the present paper differs from Vulanović (to appear a, b) by the inclusion of all mathematically possible PoS systems, which include the HRS types. This paper is more theoretical – the HRS sample of languages is not considered, nor are grammatical markers that can be found in natural languages.

2. Possible PoS Systems

Table 1 is based on the following hierarchy (HRS):

$$P > R > r > p. \quad (2)$$

The original meaning of (2) is that if a propositional function is positioned more to the left, languages are more likely to have a separate word class for this function. However, Hengeveld & van Lier (to appear) – from now on HvL – report that the Tibeto-Burman language Garo does not conform to the hierarchy in (2) because it has manner adverbs but not adjectives (this seems to be a reclassification of Garo which is described in HRS as an intermediate type language between HRS5 and HRS6). Therefore, (2) can be replaced with

$$P > R > \{r, p\}. \quad (3)$$

I read (3) the way it looks, without attaching word classes to its meaning. Thus, (3) simply tells me that if a language has a propositional function, it also has all propositional functions to the left. Table 3 shows what combinations of propositional functions are possible under (3). There are two possibilities with three propositional functions. Mathematically speaking, they are equivalent because both have two heads and one modifier.

Table 3
Possible combinations of propositional functions when (3) is assumed

n	P	R	r	p
4	+	+	+	+
3	+	+	+	-
3	+	+	-	+
2	+	+	-	-
1	+	-	-	-

When searching for theoretically possible PoS systems, I restrict myself to the combinations presented in Table 3. My PoS system types are given in Table 4. The table also shows what types are attested in natural languages according to HRS and HvL. Types 1.3 and 3.3 comprise two subtypes which are mathematically equivalent. The subtypes of 2.4, 3.4, and 2.3, referred to in Table 4, are not mathematically equivalent – they are essentially different. Those three PoS types are all flexible. This means that at least one word class performs more than one propositional function. In all three types it is possible to have one word class functioning as both heads – I label this subtype HH. It is also possible to have one word class for the head and modifier of the same phrase (subtype HM) and one word class for the head of one phrase and the modifier of the other phrase (subtype H×M). Types 2.4 and 3.4 also have an MM subtype, in which there is one word class with both modifier functions. Within type 2.4, there are additionally two subtypes in which one word class has three functions and the other word class only has one function. I denote these subtypes according to the uniquely assigned word class – it can be assigned either to a head

(subtype 2.4H) or to a modifier (subtype 2.4M). Most of the described subtypes have further subtypes, but those are mathematically equivalent. Tables 5-7 present all the subtypes. The subtypes containing attested languages are boldfaced and this is continued throughout the rest of the paper.

Table 4
PoS system types ($^+$ indicates types or subtypes not covered by the HRS classification)

type	P	R	r	p	attested in
1.4	C	C	C	C	HRS
2.4	5 subtypes				see Table 5
3.4	4 subtypes				see Table 6
4.4	V	N	a	m	HRS
1.3r$^+$	C	C	C	-	HRS
1.3p$^+$	C	C	-	C	-
2.3$^+$	3 subtypes				-
3.3r	V	N	a	-	HRS
3.3p$^+$	V	N	-	m	HvL
1.2$^+$	C	C	-	-	-
2.2	V	N	-	-	HRS
1.1	V	-	-	-	-

Table 5
Subtypes of PoS system type 2.4

2.4 subtypes		P	R	r	p	attested in
H	**P** (HRS2)	V	Λ	Λ	Λ	HRS
	R	X	N	X	X	-
M	r	X	X	a	X	-
	p	X	X	X	m	-
HH = MM		H	H	M	M	-
HM		$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{P}$	HvL
H×M		X	Y	X	Y	-

Table 6
Subtypes of PoS system type 3.4

3.4 subtypes		P	R	r	p	attested in
HH		H	H	a	m	-
HM	Pp	$\mathcal{P}$	N	a	$\mathcal{P}$	-
	Rr	V	$\mathcal{N}$	$\mathcal{N}$	m	HvL
H×M	Pr	X	N	X	m	-
	Rp	V	X	a	X	-
MM (HRS3)		V	N	M	M	HRS

Table 7
Subtypes of PoS system type 2.3, none attested

2.3 subtypes		P	R	r	p
HH	HHr	H	H	a	-
	HHp	H	H	-	m
HM	HMr	V	$\mathcal{N}$	$\mathcal{N}$	-
	HMp	$\mathcal{P}$	N	-	$\mathcal{P}$
H×M	H×Mr	X	N	X	-
	H×Mp	V	Λ	-	Λ

Tables 5-7 make use of some word classes that have not been mentioned thus far. $\mathcal{P}$ and $\mathcal{N}$ denote respectively the classes of *predicatives* and *nominals* (HvL). X and Y are just formal labels in the absence of meaningful terminology and notation. X is used generically – it is not the same word class in different subtypes. Finally, note the dual meaning of H and M. On the one hand, H stands for either head (H = P or H = R) and M for either modifier function (M = r or M = p). On the other hand, they also denote word classes (H is introduced in (1) and M in Table 1). The intended meaning of H and M can be concluded from the context.

It is explicitly stated in HvL that subtypes 3.4Pp and 3.4HH are impossible under the principles proposed in that paper. Their absence in natural languages is taken as a confirmation of those principles. The other unattested subtypes are not even mentioned – they must look very awkward to linguists. However, they are all theoretically (mathematically, logically) possible and this is why they are included here. By considering unattested structures, we may be able to get a better insight into the attested ones. One of the questions that can be raised is whether grammar efficiency plays a role in determining what kind of PoS systems occur in natural languages. I turn to this question in next two sections.

3. Grammar Efficiency

Grammar efficiency is a concept introduced in Vulanović (1991) and further developed in Vulanović (1993, 2003, 2007). The most general formula defining grammar efficiency, *Eff*, is given below:

$$Eff = \gamma\, Q\, n/k. \qquad (4)$$

In this formula, γ is a scaling coefficient ensuring that *Eff* = 1 for maximally efficient grammars. *n* is a measure of the amount of linguistic information that needs to be conveyed and *k* is the number of grammatical devices used to convey this information. Formula (4) shows that *Eff* is greater if the grammar conveys more information with fewer conveyors. In this paper, according to the already introduced notation, *n* is the number of propositional functions and *k* is the number of word classes. The quantity *Q* represents the parsing ratio; its precise definition is given below.

Both γ and the fraction *n*/*k* are needed in (4) to enable comparisons of grammars belonging to different classes, that is, grammars with different *n* or *k*. Such comparisons are not made in the present paper. Grammar efficiency is used here to compare the essential subtypes within their

corresponding PoS type (2.4, 3.4, or 2.3). Therefore, the parsing ratio *Q* suffices for this kind of comparison and is the only quantity to be calculated here. *Q* is some sort of absolute grammar efficiency (cf. absolute grammar complexity in Vulanović (2007)), as opposed to (4), which means relative grammar efficiency.

The parsing ratio is related to the way sentences permitted in the grammar are parsed. Only simple intransitive sentences are considered here. The information they have to convey is represented by any of the following four sets:

{P, R}, {P, R, r}, {P, R, p}, and {P, R, r, p}.

There are 38 possible ways of ordering the elements of the above sets (2! + 3! + 3! + 4! = 38). However, discontinuous phrases are not going to be considered in this paper. This leaves the following 18 orders of propositional functions:

PR, RP, PpR, pPR, RPp, RpP, PRr, PrR, RrP, rRP, (5a)

PpRr, PpRr, pPRr, pPrR, RrPp, RrpP, rRPp, rRpP. (5b)

Example 1. As an illustration, consider now the attested **2.4P** subtype (HRS2). According to Table 5, the mapping, denoted below by Φ, between word classes and propositional functions in this subtype is as follows:

Φ: $V \rightarrow P, \quad \Lambda \rightarrow R, r, p.$ (6)

Suppose we want to describe now a **2.4P** language in which the propositional functions appear ordered as in RrPp. Then, there are four possible intransitive sentences (strings of word classes): ΛV, ΛΛV, ΛVΛ, and ΛΛVΛ. They are parsed as follows:

$\Lambda V \rightarrow RP, \quad \Lambda\Lambda V \rightarrow RrP, \quad \Lambda V\Lambda \rightarrow RPp, \quad \Lambda\Lambda V\Lambda \rightarrow RrPp.$ (7)

No sentence has ambiguous interpretation. Let ρ indicate the number of successful parses of all permitted sentences and ρ_0 the number of ambiguous parses. Then, according to (7), $\rho = 4$ and $\rho_0 = 0$ in this example.

However, the mapping in (6) can allow for more than four sentences, i.e. the order of propositional functions does not have to be fixed as initially assumed. There are 9 permutations of the four sentences in (7):

ΛV, VΛ, ΛΛV, ΛVΛ, VΛΛ, ΛΛΛV, ΛΛVΛ, ΛVΛΛ, VΛΛΛ. (8)

Each of them can be interpreted in a unique way. VΛΛΛ, for instance, can be parsed as either PpRr or PprR (recall that PPh and RPh have to be continuous) but then a word-order rule can be imposed to select one of the two possible parses for the meaning of this sentence. Let ρ_A be the greatest possible number of unambiguous parses for the given mapping Φ, which assigns word classes to propositional functions. Thus, in this example, $\rho_A = 9$. Note that ρ_A cannot be greater than 18 because of (5).

If it is assumed that parsing is done from left to right, one word at a time, and the number of words is not known in advance, then there are more parsing attempts than the successful parses. Under these assumptions, there are three parsing attempts to parse ΛV for instance. One of the attempts can be finished successfully, giving RP, but Λ can be initially also interpreted as either r or p. These two parsing attempts are abandoned when it is realized that r is not followed by R and, respectively, when it is realized that there is no other word after V. The parsing attempts, successful or not, are represented in this case as follows:

ΛV → RP/r-/pP-.

The symbol '-' indicates here that the parse is abandoned. Let ρ* denote the number of parsing attempts, of which at least one is successful, applied to all permutations of all possible sentences. There are in all ρ* = 34 parsing attempts of the sentences in (8), the above three and the following 31:

VΛ → PR/Pr-/Pp-, ΛΛV → RrP/RpP/rRP/p-,

ΛVΛ → RPp/r-/pPR/pPr-, VΛΛ → PRr/PrR/PpR/Ppr-,

ΛΛΛV → RrpP/Rp-/rRpP/p-, ΛΛVΛ → RrPp/RpP-/rRPp/p-,

ΛVΛΛ → RPp-/r-/pPRr/pPrR, VΛΛΛ → PRr-/PrR-/PpRr/PprR.

Example 2. Assume now that word order in **2.4P** is defined by RrpP. Then there are just three possible sentences: ΛV, ΛΛV, and ΛΛΛV. ΛΛV is ambiguous because of

ΛΛV → RrP/RpP.

This ambiguity cannot be avoided since both RrP and RpP have to be conveyed and no other sentence can do that. The number of successful parses of all three sentences is still 4 (ρ = 4), but now $\rho_0 = 2$. The values of ρ_A and ρ* remain the same as in Example 1.

The different parsing-related counts, introduced above, are used to define the parsing ratio, cf. Vulanović (to appear b),

$$Q = (\rho - \rho_0)/\rho^*. \qquad (9)$$

Formula (9) is motivated by the following:

- The grammar is more efficient if it has fewer word order rules, that is, if ρ is greater.
- The grammar is more efficient if it permits fewer ambiguous sentences, that is, if ρ_0 is less.
- The grammar is more efficient if its permitted sentences require fewer parsing attempts, that is, if ρ* is less.

ρ_A is not used in (9) but it is part of another ratio,

$$Q_A = \rho_A/\rho^*,$$

introduced in Vulanović (to appear b). Since $0 \leq \rho - \rho_0 \leq \rho_A \leq \rho^*$, it follows that

$$0 \leq Q \leq Q_A \leq 1.$$

$Q = Q_A$ if $\rho = \rho_A$ while $\rho_0 = 0$. If the assignment (mapping Φ) of word classes to propositional functions is kept fixed and word-order rules are varied, Q changes its value and Q_A is the maximum that Q attains. For this reason, max Q will be written instead of Q_A from this point on. If assignment (6) is assumed, the greatest possible value of Q is max $Q = 9/34$ because $\rho_A = 9$ is calculated in Example 1. The PoS systems in Examples 1 and 2 have fixed word order and their parsing ratios are less than this maximum: $Q = (4 – 0)/34 = 2/17$ in Example 1 and $Q = (4 – 2)/34 = 1/17$ in Example 2.

In general, ρ^* depends on how much mapping Φ differs from a one-to-one mapping. If Φ is one-to-one, then $\rho^* = \rho_A$. This is the case whenever the PoS system is rigid ($k = n$) because then every sentence requires just one parsing attempt, which is successful. If $\rho^* = \rho_A$, the greatest possible value of the parsing ratio is 1. This is achieved by making word order completely free, so that $\rho = \rho_A$ (see Example 3 below). Otherwise, if $k < n$, word order has to be restricted in order to enable unambiguous parsing. Then, $\rho^* > \rho_A$ and $Q < 1$. If there are more word-order rules, the total number of grammatical rules increases and grammar efficiency and Q decrease. By considering all permutations of all sentences when calculating ρ_A and ρ^*, we establish how restricted the actual word order is.

Example 3. Consider a PoS system of type **4.4** (HRS4). This is where mapping Φ is one-to-one, see Table 1. If word order is assumed fixed, say VmNa, then ρ is just 4 because the following parses are the only possible ones:

$$VN \rightarrow PR, \quad VmN \rightarrow PpR, \quad VNa \rightarrow PRr, \quad VmNa \rightarrow PpRr.$$

However, every permutation of every sentence can be parsed unambiguously (it is impossible to create ambiguity in this case) and in a single parsing attempt (for instance, aNVm → rRPp). Therefore, $\rho_A = 18$ (because of (5)) and ρ^* has the same value. This is why max $Q = 1$ and the minimum Q value is min $Q = 4/18=2/9$.

Example 4. In order to illustrate the procedure for finding the parsing ratio once more, let us consider PoS system type 2.4HH. In this type, mapping Φ is like in (1), which is repeated here for convenience:

$$\Phi: \quad H \rightarrow P, R, \quad M \rightarrow r, p.$$

The greatest possible number of unambiguous sentences is $\rho_A = 8$. Those sentences are given below together with their parsing attempts:

$$HH \rightarrow PR/RP, \quad HHM \rightarrow PRr/RPp,$$

$$HMH \rightarrow PrR/PpR/RrP/RpP, \quad MHH \rightarrow rRP/pPR,$$

HMHM → PrR-/PpRr/RrPp/RpP-, HMMH → Pr-/PprR/RrpP/Rp-,

MHHM → rRPp/pPRr, MHMH → rRpP/pPrR.

The total of parsing attempts is ρ = 22. Therefore, the greatest possible value of the parsing ratio for type 2.4HH is max Q = 8/22 = 4/11.

In order to find min Q, all possible fixed word orders have to be explored. If there is no ambiguity, the smallest value of ρ, ρ = 4, gives a good candidate for the smallest possible value of Q. However, if there is ambiguity, then $\rho - \rho_0$ is less than 4. Therefore, it should be analyzed if there is a fixed word order which produces ambiguity. This analysis is simplified by the fact that, generally speaking, when n = 4, only three-word sentences can be ambiguous. In type 2.4HH, there is one fixed order or propositional functions, PprR, in which sentence HMH is ambiguous:

HMH → PrR/PpR.

This shows that min Q = (4 – 2)/22 = 1/11 in this PoS system type.

4. Results and Conclusions

PoS system types 2.4, 3.4, and 2.3 are considered in this section. All subtypes within each type are compared using their parsing-ratio values. The whole range of values is of interest, the endpoints being min Q and max Q. The calculations are done like in the examples of the previous section. Only final results are reported here – see Table 8 – and details are omitted. In the case of subtypes having two variants (**2.4H**, 2.4M, **3.4HM**, 3.4H×M, and all 2.3 subtypes), the variants share the same numbers. This is one manifestation of their mathematical equivalence.

Table 8
ρ_A, ρ*, max Q, min Q, and Q_0 for all
subtypes of PoS system types 2.4, 3.4, and 2.3.

subtype	ρ_A	ρ*	max Q	min Q	Q_0
2.4H	9	34	.265	.059	.118
2.4M	9	37	.243	.108	
2.4HH	8	22	.364	.091	.182
2.4HM	8	28	.286	.143	
2.4H×M	12	34	.353	.118	
3.4HH	15	24	.625	.167	
3.4HM	12	24	.500	.167	
3.4H×M	17	30	.567	.133	
3.4MM	16	28	.571	.071	.143
2.3HH	4	8	.500	.286	
2.3HM	4	7	.571	.250	
2.3H×M	5	10	.500	.200	

In all calculations of min Q, $\rho = 4$ is used when $n = 4$ and $\rho = 2$ when $n = 3$. Ambiguity is possible when word order is fixed in subtypes **2.4H**, 2.4HH, and **3.4MM** – this is why their min Q values are lower. In those three subtypes, there is a fixed word order with an ambiguous sentence ($\rho_0 = 2$). This is shown for **2.4H** and 2.4HH in Examples 2 and 4 respectively. As for **3.4MM**, the PprR order permits an ambiguous three-word sentence: VMN → PrR/PpR. At the same time, the three subtypes have other fixed word orders which do not produce ambiguity. The value of the parsing ratio in this case is denoted in Table 8 by Q_0. In the other subtypes, which have no ambiguity, $Q_0 = \min Q$.

A graphical representation of these results is given in Figures 1-3 for each PoS system type separately. The span of Q values is shown as a vertical line-segment and the midpoint of each segment (the midrange = average of max Q and min Q) is indicated. Each figure is accompanied with a table showing how the subtypes are ranked within their own type.

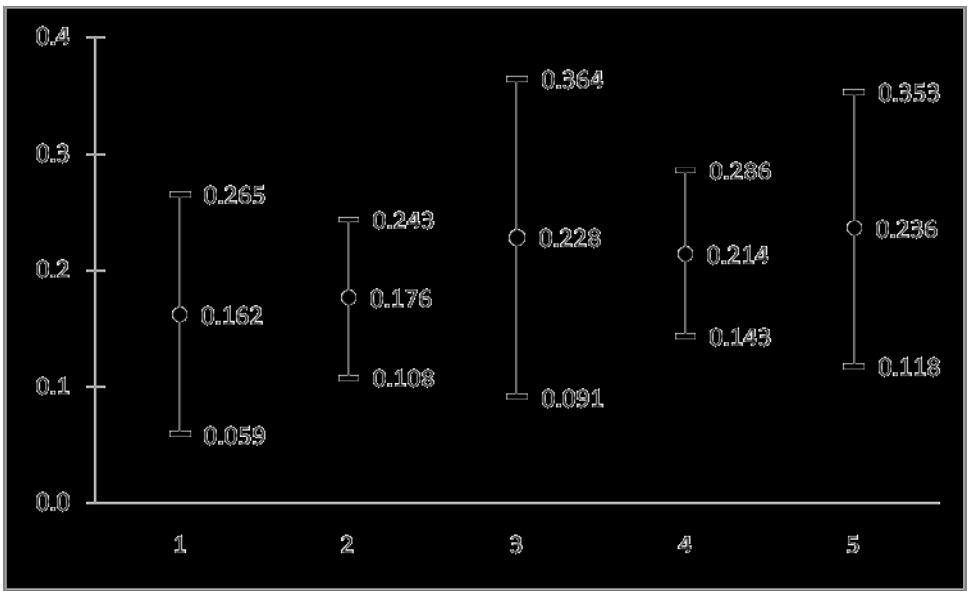


Fig. 1. Parsing ratio for 2.4 subtypes:
1 - **2.4H**, 2 - 2.4M, 3 - 2.4HH, 4 - **2.4HM**, 5 - 2.4H×M

Table 9
2.4 subtypes ranked with respect to their parsing-ratio values

rank	max Q	min Q	midrange	Q_0
1	HH	**HM**	H×M	HH
2	H×M	H×M	HH	**HM**
3	**HM**	M	**HM**	**H**, H×M
4	**H**	HH	M	-
5	M	**H**	**H**	M

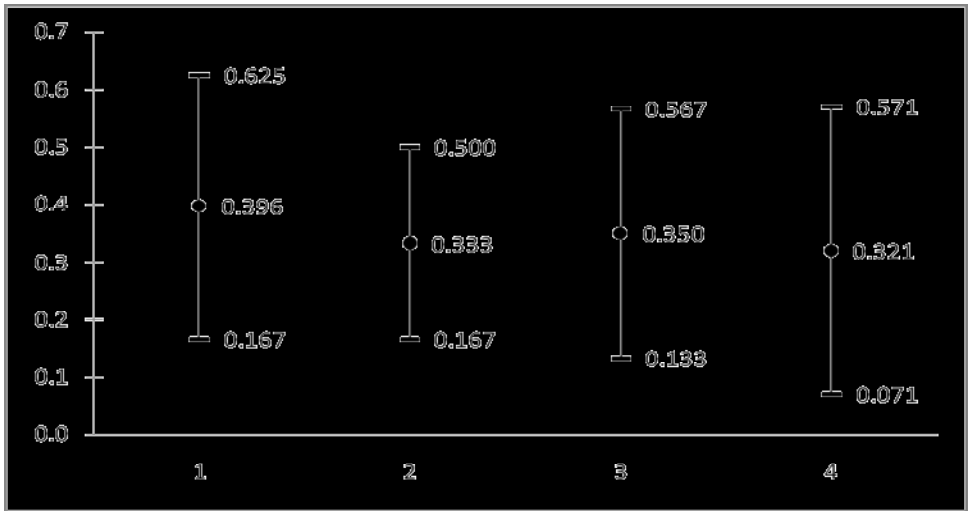


Fig. 2. Parsing ratio for 3.4 subtypes:
1 - 3.4HH, 2 - **3.4HM**, 3 - 3.4H×M, 4 - **3.4MM**.

Table 10
3.4 subtypes ranked with respect to
their parsing-ratio values.

rank	max Q	min Q	midrange	Q_0
1	HH	HH, **HM**	HH	HH, **HM**
2	**MM**	-	H×M	-
3	H×M	H×M	**HM**	**MM**
4	**HM**	**MM**	**MM**	H×M

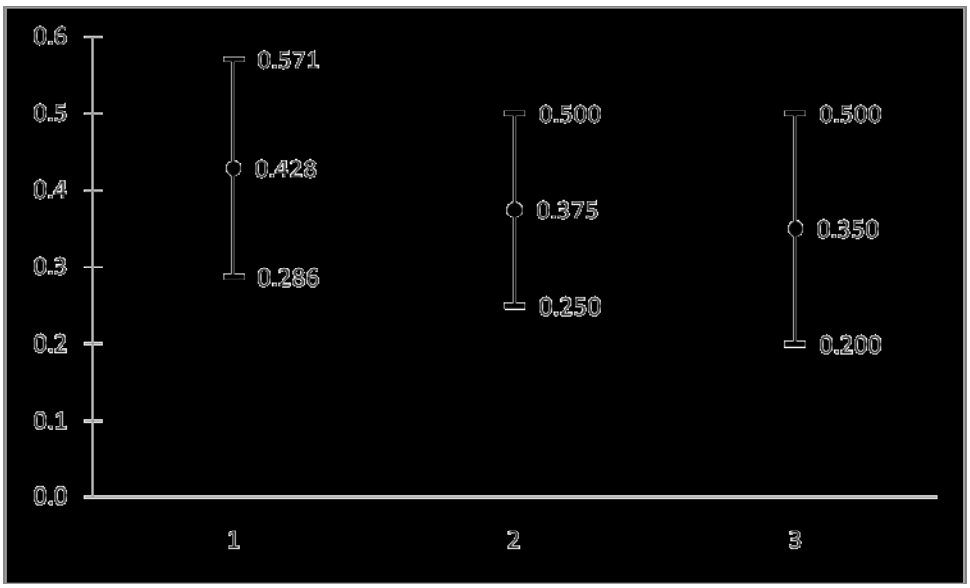


Fig. 3. Parsing ratio for 2.3 subtypes:
1 - 2.3HH, 2 - 2.3HM, 3 - 2.3H×M

Table 11
2.3 subtypes ranked with respect to
their parsing-ratio values

rank	max *Q*	min *Q*	midrange
1	HH	HH	HH
2	HM, H×M	HM	HM
3	-	H×M	H×M

There are several conclusions that can be derived from the presented results, particularly from Tables 9-11. The greatest possible value of the parsing ratio, and, therefore, of grammar efficiency, belongs invariably to structures with maximally free word order and the unattested HH PoS system type. Within types 3.4 and 2.3, HH holds the top rank even when word order is fixed. That this is not so in type 2.4, is because of ambiguity which is possible in some fixed word orders. Other fixed word orders, which do not produce ambiguity, place 2.4HH at the top rank again.

PoS system subtypes containing attested structures are ranked lower when word order is as free as possible. There are two such subtypes within each 2.4 and 3.4 type and fixed word order has opposite effects on them. Efficiency of both **2.4HM** and **3.4HM** increases and reaches the top rank of the min *Q* values. At the same time, efficiency of **2.4H** and **3.4MM** decreases to the lowest rank. This is again because of possible ambiguous sentences. The values of Q_0 show that both **2.4H** and **3.4MM** are ranked better when only unambiguous fixed word orders are considered. It is interesting to point out that **2.4H** (which contains HRS2) and **3.4MM** (= HRS3) represent structures which are more typical and frequent in natural languages than those belonging to subtypes **2.4HM** and **3.4HM** (these have been just recently reported in HvL). Therefore, it can be stated that the results presented here indicate that attaining the highest possible grammar efficiency (parsing ratio) is not what languages try to achieve during their evolution. If, according to functional linguistics, languages are shaped by usage, then grammar efficiency is not one of the shaping factors. This is in agreement with the findings in Vulanović (to appear a, b).

Table 11 contains no attested structures, but subtype 2.3HM is close to the attested **2.4HM** and **3.4HM**, whereas no HH or H×M type is attested. Type 2.3 is simpler than types 2.4 and 3.4, and the results for 2.3 are more consistent: the top and bottom ranks are occupied by subtypes which are unlikely to be ever found in natural languages, while the more plausible type 2.3HM is ranked in the middle.

This paper presents a mathematical approach to the classification of PoS systems. Considered here are all PoS systems which are possible under two assumptions: assumption (3) and the assumption on the continuity of predicate and referential phrases (5). This is regardless of whether the system is attested or not, which is contrary to the linguistic approach. Linguists look into natural languages and then try to classify them according to some features. If a new type is discovered later on, the classification and the underlying explanations may have to be changed. This is indeed what is currently going on with the classification of PoS system types, witness HRS and HvL. On the other hand, theoretical considerations may require that some unattested types be analyzed. For instance, I had to consider the subtypes which are labeled here **2.4HM** and **3.4Rr** even before I was aware of HvL. I needed them for grammar-efficiency calculations in

Vulanović (to appear b). In general, if linguistic classification is approached so that all theoretical possibilities are covered, then, although it is necessary to consider abstract structures, the result already contains types which may be attested in the future.

Together with Vulanović (to appear b), this paper shows that different versions of grammar efficiency (including the parsing ratio) can be used to compare various PoS systems quantitatively. At this stage, grammar efficiency is just a measure based on a mathematical model. Further empirical study, involving data from typological samples of as many languages as possible, is needed in order to establish what kind of dependencies exist between this measure and other properties of language.

References

Baker, M. C. (2003). *Lexical Categories.* Cambridge: Cambridge University Press.

Croft, W. (2000). Parts of speech as language universals and as language-particular categories. In: Vogel, P. M., & Comrie, B. (Eds.), *Approaches to the Typology of Word Classes (Empirical Approaches to Language Typology 23): 65-102.* Berlin/New York: Mouton de Gruyter.

Göksel, A., & Kerslake, C. (2005). *Turkish. A Comprehensive Grammar.* New York: Routledge.

Hengeveld, K. (1992). Parts of speech. In: Fortescue, M., Harder, P., & Kristoffersen, L. (Eds.), *Layered Structure and Reference in Functional Perspective: 29–55*. Amsterdam/ Philadelphia: John Benjamins.

Hengeveld, K., & Rijkhoff, J. (2005). Mundari as a flexible language. *Linguistic Typology 9, 406–431*.

Hengeveld, K., Rijkhoff, J., & Siewierska, A. (2005). Parts-of-speech systems and word order. *Journal of Linguistics 40, 527–570*.

Hengeveld, K., & van Lier, E. (to appear). An implicational map of parts of speech.

Newmeyer, F. J. (2005). *Possible and Probable Languages.* Oxford: Oxford University Press.

Rijkhoff, J. (2007). Word classes. *Language and Linguistics Compass 1, 709–726.*

Vulanović, R. (1991). On measuring grammar efficiency and redundancy. *Linguistic Analysis 21, 201–211*.

Vulanović, R. (1993). Word order and grammar efficiency. *Theoretical Linguistics 19, 201–222*.

Vulanović, R. (2003). Grammar efficiency and complexity. *Grammars 6, 127–144.*

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics 20, 399-427*.

Vulanović, R. (to appear a). The combinatorics of word order in flexible parts-of-speech systems. *Glottotheory.*

Vulanović, R. (to appear b). Efficiency of flexible parts-of-speech systems. Paper presented at the 5th Trier Symposium on Quantitative Linguistics, Trier, December 6-8, 2007.

Glottometrics 17, 2008, 66-78

On meaning diversification in English

Fan Fengxiang, Dalian[1]
Gabriel Altmann, Lüdenscheid

Abstract. Individual meanings (senses) of words have different frequencies of occurrence. They form rank-frequency distributions with different properties. To capture these properties the Shenton-Skees-geometric distribution and the Popescu-formula are used. Semantic concentration is determined by a concentration measure and expressed by Ord´s criterion.

Keywords: Diversification, English, semantic concentration, Shenton-Skees-geometric distribution, Zipf´s law, Ord´s criterion

1. Introduction

Polysemy considered in isolation has four main quantitative aspects: (1) The distribution of words (f_x) containing x meanings in the lexicon, representing a frequency spectrum, usually called Krylov´s law. (2) The distribution of frequencies of individual meanings (senses) of an individual word in texts, representing a rank-frequency distribution usually called Beöthy´s law. (3) The relationship between the frequency of a word and the number of its meanings. Such relationship is expressible by a function (Zipf 1945, 1949; Levickij 2005). (4) The number of word classes in which a word can penetrate – either without any change, as is usual with the analytic way of expression (e.g. English: *the hand, to hand*) or adopting affixes (e.g. German: *Bild, Vorbild, bilden, bildhaft, vorbildlich, gebildet*).

There have been a number of models attempting to account for all the four aspects. However, the research in the area of meaning diversification is by no means complete due to the following factors: (i) There are manifold causes leading to diversification (e.g. random fluctuation, environmentally conditioned variation, conscious change, self-regulative triggering, system modification, fulfilling of Köhlerian requirements, metaphorization, cf. Altmann 2005). One can never ascertain which causes, or better, factors can account for the actual anomaly. The majority of models have a common origin but in the course of research the argumentation is continuously widened. Beginning with the simple proportionality approach one tries to capture the rank-order of diversification by modifications, generalizations according to Feller, partial sums distributions, compounding, mixing, Leopold´s (1998) recurrence function and stochastic processes, yielding an extensive battery of models, but the interaction of factors can be manifold that no theory will ever be able to capture and to incorporate them directly in the argumentation. (ii) No model holds true for all data. This can be caused not only by the above mentioned circumstances but also by differently sampled data or "wrong" data, a special development of polysemy in different word classes, or even by a given language displaying special boundary conditions. This in turn evokes two questions: (a) what is the mechanism behind diversification and (b) how can we reconcile the methodological striving for simple functions (e.g. those with few parameters) with the existence of an enormous number of different boundary conditions accompanying every word and its semantic

[1] Address correspondence to: fanfengxiang@yahoo.com

diversification? If there is a unique mechanism, we need a function with a sufficient number of parameters expressing the boundary conditions; but if the diversification of a word is not advanced, we cannot estimate many parameters from "short" data, i.e. a complex model must have different special and/or limiting cases.

Thus there may be different strategies of approaching the problem of meaning diversification, but there is no independent criterion which would allow us to make decisions about the choice of the "correct" model. Peculiar enough, from the deductive point of view the systematization of the resulting formulas in a broader theory would be sufficient even if exceptions from good fitting would occur; however, a mere good fitting of an ad hoc formula, i.e., a good result from an inductive procedure could never satisfy our requirements because it excludes a possible explanation, and the fitting itself using the chi-square criterion is usually a very shaky background. The weaknesses of the goodness-of-fit procedures are generally known, whether one approaches the problem deductively or inductively.

There are different ways out of this dilemma. First, one need not use a probability mass function for modelling; a simple sequence (i.e. non normalized) would do. Second, it need not be a discrete model; it may be a continuous function or density. The two types of modelling are merely approximations both from the empirical and theoretical points of view. Empirically, meaning is a fuzzy entity, for the capture of which we use discrete or continuous models; such models are our conceptual constructs, nothing else. Theoretically, continuous and discrete models are mutually transformable in one another (cf. Mačutek 2007; Mačutek, Altmann 2007) and some very general models have been set up for both variants (cf. Wimmer, Altmann 2005). Third, testing with a chi-square criterion is problematic both with very small and very large samples, and the largest deviations occur usually in non-frequent, i.e. irrelevant classes in which the asymmetry of observed and expected frequencies causes an additional negative weight. And since the rejection criteria (significance levels) are not objective entities, one can perform goodness-of-fit tests also using e.g. the determination coefficient, which is free of degrees of freedom and sample size.

Now, in order to obtain the frequencies of individual senses of a word, it is sufficient to identify a sense as different from all the other senses. Thus by ordering the frequencies of individual senses one obtains a discrete rank-frequency distribution. The modelling may proceed using a function, a density, a sequence or a probability mass function. As a matter of fact, all these approaches have been tried and we shall show some of them.

In English only the diversification of the preposition *in* has been studied by Hennern (1991) using different models. Here we proceeded differently. We randomly chose 165 words from a dictionary and used the frequencies of individual senses contained in WordNet of the Princeton University (http://www.cogsci.princeton.edu/~wn/). The beginning of our list is as follows:

Word	Word class	Sense number	Frequency
Animal	N	1	67
Animal	Adj	2	1, 1
Ash	N	3	2, 1, 1
Ash	V	1	1
Back	N	9	53, 12, 4, 1, 1, 1, 1, 1, 1
Back	V	10	7, 6, 4, 4, 2, 1, 1, 1, 1, 1
Back	Adj	3	15, 1, 1
Back	Adv	6	92, 36, 24, 15, 14, 1

A word can appear in several word classes and continue to diversify semantically in each class. The reading of the data is as follows: *Animal* appears in 2 classes (N, Adj), has 3

meanings (1+2) and its rank-frequency distribution is 67, 1, 1. We pool the classes, order the frequencies and obtain the following results:

Word	**Class number C_k**	**Sense number S_k**	**Distribution $\{f_i\}$, $i = 1,2,\ldots,S_k$**
Animal	2	3	67,1,1
Ash	2	4	2,1,1,1
Back	4	28	92,53,36,24,15,15,14,12,7,6,4,4,2,1,1,1,1,1,1,1,1,1,1,1,1,1.

The frequencies mean the number of their occurrences in the source texts. As can be seen, diversification can display three different features: (a) Different number of classes to which a word belongs (*C*), (b) different number of senses (*S*), (c) different degrees of concentration on the main sense signalizing the extent of diversification. For example, *Animal* appears in two classes and has 3 senses but with a strong concentration on one of them; *Ash* has four meanings but the concentration is smaller: the main meaning is not especially preferred. *Back* is strongly diversified in all three aspects.

2. Characterization

2.1. Expansion in classes

Considering aspect (4) mentioned at the beginning, if there are word classes in a language, then originally, i.e. in the moment of its creation, every word belonged to one class only. Then step by step some words expanded to other classes. This expansion, which is very slow, can be captured by a simple geometric series. But as time goes on and the given language develops to an analytic type, the expansion may accelerate and arrive at a state in which the expansion is not any more monotonously decreasing but gets a concave shape. The number of words belonging to two classes gets greater than that of words belonging to one class only. In this situation one either changes the modelling strategy and considers the problem from the synchronic point of view, or adheres to the historical view and modifies the geometric distribution adequately.

In English the penetration of words in different classes is very common. Our random sample yielded the numbers presented in Table 1, where x is the number of classes to which the words belong and f_x the number of words belonging to x classes. Following the historical perspective we performed the Gram-Charlier expansion of the geometric distribution and obtained the Shenton-Skees-geometric distribution (cf. Shenton, Skees 1970; Wimmer, Altmann 1999)

$$(1) \qquad P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right], \quad x = 1,2,3,\ldots$$

where a and p are parameters, $0 < p < 1$, $q = 1\text{-}p$, $0 < a < 1/q\text{-}1$. As can be seen in Table 1 and Figure 1, the fitting is almost perfect. In linguistics, this distribution is used with script diversification problems having the same historical perspective (cf. Mačutek 2008).

Table 1
Fitting the Shenton-Skees-geometric distribution
to class expansion of English words

x	f_x	NP_x
1	25	25.00
2	110	110.00
3	25	24.95
4	5	5.05
Sum	165	
p = 0.8851, a = 6.3839, X^2 = 0.0005, DF = 1, P = 0.98		

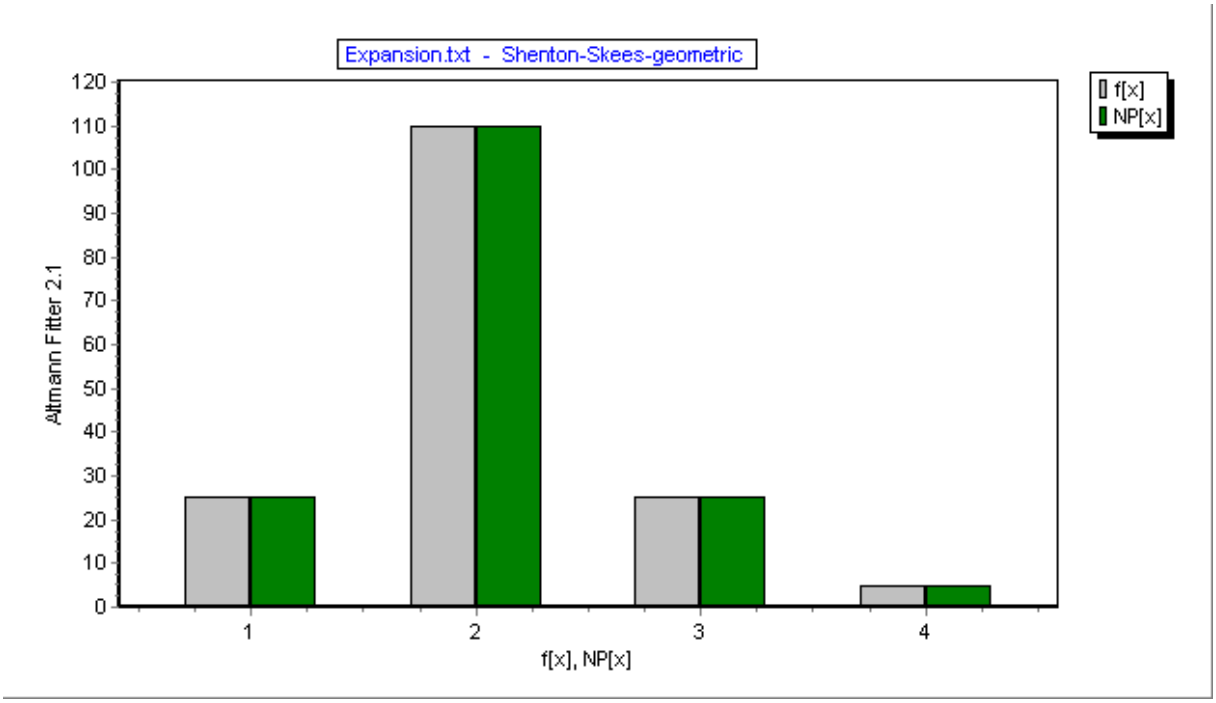


Figure 1. Fitting the Shenton-Skees-geometric distribution
to class expansion of English words

From the synchronic point of view perhaps a Poisson process would be adequate, but we shall not touch the problem until the words from an entire English dictionary have been processed.

2.2. Sense diversification

Here, our main aim is the study of individual semantic diversification of words. According to Zipf (1935, 1949) the speaker tends to express everything with only one word – a technique saving both coding and production effort. However, this would cause maximal decoding effort from the part of the hearer for whom "one word, one meaning" would be the best solution. Hence both must try to find a compromise affording to both parties involved in communication the minimum of effort. Since the speaker is the active party, there will be at least some words having several meanings. However, the decoding may be facilitated by specific

environment of the word, fixed phraseology or even the concentration of occurrences to one of the senses. Strong concentration allows us to guess the "proper" sense with high probability.

In order to measure the concentration of occurrences, we use the Repeat rate *R* (or Herfindahl´s concentration measure) defined as

$$R = \frac{1}{N^2}\sum_{i=1}^{S} f_i^2 . \qquad (2)$$

where *N* is the sum of the frequencies, S is the number of senses. For *Animal* we obtain

$$R = [67^2 + 1^2 + 1^2]/69^2 = 0.9433$$

expressing a very high concentration. For *Ash* we obtain *R* = 0.2800 and for *Back R* = 0.1584. Thus the main sense of *Animal* can be guessed with high probability, but the specific meaning of *Back* could be guessed with difficulty though the general meaning is unambiguous. In Table 2 the repeat rates of all the 165 words are presented.

Table 2
Repeat rates (meaning concentration) of 165 English words

Word	Classes	R	Word	Classes	R	Word	Classes	R
Animal	2	0.9433	Full	4	0.3234	Salt	3	0.4661
Ash	2	0.2800	Give	2	0.1242	Sand	2	0.5408
Back	4	0.1584	Good	3	0.4056	Say	2	0.5769
Bad	3	0.5085	Grass	2	0.6760	Scratch	2	0.1867
Bark	2	0.1667	Green	3	0.3740	Sea	2	0.7869
Belly	2	0.3674	Guts	3	0.1875	See	2	0.2899
Bird	2	0.7454	Hair	1	0.8511	Seed	2	0.2296
Bite	2	0.2544	Hand	2	0.6742	Sew	1	0.7222
Black	3	0.3885	Head	2	0.3970	Sharp	3	0.1121
Blood	2	0.8867	Hear	1	0.6268	Sing	1	0.4532
Blow	2	0.1493	Heart	1	0.3233	Sit	1	0.5486
Bone	3	0.3910	Hit	2	0.1578	Skin	2	0.2015
Breast	2	0.3056	Hold	2	0.1365	Sky	2	0.9608
Breath	1	0.5813	Horn	2	0.1856	Sleep	2	0.5394
Burn	2	0.1225	Hunt	2	0.0914	Smell	2	0.2136
Child	1	0.6279	Husband	2	0.9722	Smoke	2	0.5786
Cloud	2	0.3241	Ice	2	0.6062	Smooth	3	0.2089
Cold	2	0.3255	Kill	2	0.7262	Snake	2	0.4935
Come	2	0.2492	Knee	1	0.8631	Snow	2	0.2871
Corrrect	2	0.2012	Know	2	0.4198	Spit	2	0.3951
Count	2	0.2389	Lake	1	0.4400	Split	3	0.0780
Cut	3	0.6173	Laugh	2	0.6507	Squeeze	2	0.1437
Day	1	0.3495	Leaf	2	0.6480	Stab	2	0.1667
Die	2	0.7895	Left	4	0.1457	Stand	2	0.3078

Dig	2	0.2174	Leg	1	0.7017	Star	3	0.1584
Dirty	2	0.2544	Lie	2	0.2771	Stick	2	0.0720
Dog	2	0.7096	Live	3	0.3116	Stone	3	0.3548
Drink	2	0.2838	Liver	2	0.5556	Straight	3	0.1221
Dry	3	0.1956	Louse	1	0.2800	Suck	2	0.2400
Dull	2	0.0844	Man	2	0.4482	Sun	2	0.5640
Dust	2	0.4888	Meat	1	0.5000	Swell	3	0.1319
Ear	1	0.5317	Moon	2	0.6288	Swim	2	0.7449
Earth	2	0.3685	Mother	2	0.8742	Tail	2	0.1696
Eat	1	0.5456	Mountain	1	0.8951	Think	2	0.3456
Egg	2	0.6900	Mouth	2	0.4690	Throw	2	0.2564
Eye	2	0.8257	Name	2	0.6848	Tie	2	0.1245
Fall	2	0.1120	Near	3	0.2736	Tongue	2	0.3059
Far	3	0.2617	Neck	2	0.8033	Tooth	1	0.6257
Fat	3	0.4343	New	2	0.4107	Tree	2	0.8971
Father	2	0.7052	Night	1	0.5457	Turn	2	0.2451
Fear	2	0.3913	Nose	2	0.4548	Vomit	2	0.2500
Feather	2	0.2912	Old	2	0.4090	Walk	2	0.8187
Fight	2	0.2235	Play	2	0.0922	Warm	3	0.3838
Fingernail	1	1-0000	Pull	2	0.2610	Wash	2	0.1167
Fire	2	0.3929	Push	2	0.4215	Water	2	0.5738
Fish	2	0.4380	Rain	2	0.4236	Wet	3	0.4775
Float	2	0.2140	Red	2	0.3370	White	3	0.3365
Flow	2	0.1680	Right	4	0.4100	Wife	1	1.0000
Flower	2	0.6109	River	1	1.0000	Wind	2	0.3934
Fly	3	0.2308	Road	2	0.9214	Wing	2	0.1655
Fog	2	0.5520	Root	2	0.1772	Wipe	2	0.8951
Foot	2	0.4456	Rope	2	0.4375	Woman	1	0.6951
Forest	2	0.8554	Rotten	1	0.3333	Worm	2	0.2500
Freeze	2	0.1391	Round	4	0.1076	Year	1	0.9262
Fruit	2	0.4297	Rub	2	0.6900	Yellow	3	0.4263

Peculiar enough, the number of senses is not associated with frequency. i.e. $S \neq f(N)$ or vice versa, but maybe this is only a property of our data. Though diversification of meaning can be reached only by increase of frequency, this increase does not automatically contribute to diversification.

However, the penetration of words in different word classes (C) results in decrease of concentration as expressed by the Repeat rate. Of course, this holds only on the average. If we compute the mean concentration in each of the four classes (C = 1,2,3,4), we obtain the results given in Table 3. As is usual in synergetic linguistics, many average properties are linked by a power function. Here, we obtain the relationship

(3) $R = aC^{-b} = 0.6461C^{-0.6775}$

yielding the results in the third column of Table 3. The determination coefficient is $D = 0.99$.

Table 3
Mean semantic concentration ($\overline{R}$) of words penetrating in *C* classes

C (number of classes)	$\overline{R}$ (mean concentration)	$\hat{R}$ (3)
1	0.6387	0.6461
2	0.4256	0.4040
3	0.3135	0.3069
4	0.2290	0.2526
a = 0.6461, b = 0.6775, D = 0.99		

Now, looking at the values of *R* in Table 2 we can state that 85 out of 165 words have a concentration smaller than 0.4. This means that as soon as new senses develop the main sense quickly loses its centrality and the meaning of the word develops in a state of fuzziness, a kind of multidimensional fractal. The distribution of concentrations yields a quite regular continuous form, but at the present stage of research it would be risky to make a conjecture about its form.

2.3. Ord´s indicator

If the frequencies of individual senses follow a special rank-frequency distribution, then the following circumstances must be taken into account: (1) Diversification begins with a simple development which can be modelled between the two extremes: discrete uniform distribution in which all frequencies are equal (e.g. 1,1,1,1) and the two-point distribution in which the frequencies are $f_1 = N\text{-}1$, $f_2 = 1$. Afterwards the development itself diversifies according to processes involved and according to boundary conditions which can be immensely variegated. Though a theory tends to unification, in this domain perhaps it never will be possible to combine fashion, metaphorization, reality mirrored in semantics, etc. in a unique model. (2) If one succeeds in finding a unique probabilistic model, it must have such a great number of parameters – even if they can be well interpreted – that both their estimation and the testing for goodness-of-fit will be made difficult.

In order to show that all diversifications have something in common, we compute for all words Ord´s criterion consisting of ratios of central moments (m_r): $I = m_2 / \overline{x}$ and $S = m_3/m_2$ presented for each word in Table 4. Of course, they differ strongly, but if we plot the values of <*I*,*S*> in a Cartesian coordinate system, we can state that the relationship fills a certain area. Theoretically, if the deconcentration (diversification) maximally increases, *I* increases, too, but $S = 0$; on the other hand, if concentration increases, i.e. the first value increases and the rest remains 1, one obtains $I \rightarrow 0$ and $S \rightarrow \infty$. Thus the <*I*,*S*> domain of diversification can fill the complete first quadrant. In practice it is not so, as can be seen in Figure 2. Here, evidently, the points, except for some outliers, lie around a straight line which can be expressed in the form $S = 0.7327 + 1{,}6160\ I$, yielding $D = 0.79$. It is not known which distribution encompasses the given domain. The points lie in all Ord´s domains (binomial, Poisson, negative binomial, hypergeometric, beta-Pascal, negative hypergeometric), but not all distributions are adequate as models. Some of them can be found in empirical studies and surveys (cf. Rothe 1991; Altmann 2005), others have been obtained using the FITTER software for our data.

Table 4
Ord´s coordinates <I,S> ordered according to increasing I

I	S	I	S	I	S	I	S
0.0137	0.9712	0.7276	2.5685	1.7272	2.8320	3.8178	7.2863
0.0192	0.9592	0.7384	3.3448	1.7370	5.5992	3.8592	3.7155
0.0598	1.5958	0.7570	2.3915	1.7719	9.0144	3.8592	3.7155
0.0677	1.5878	0.7773	2.5203	1.7923	6.8632	3.8963	4.3497
0.0934	0.0497	0.7773	2.5203	1.8808	3.6905	4.2372	4.1331
0.1091	1.4017	0.8065	2.7032	2.0000	1.1250	4.3874	7.9044
0.1191	0.6667	0.8214	1.8913	2.0966	2.7599	4.3896	6.1941
0.1300	1.4738	0.8333	0.0000	2.1004	2.7329	4.4089	5.7365
0.1368	2.3365	0.8355	0.9244	2.2228	4.2480	4.4544	4.9199
0.1639	1.8578	0.8426	1.6717	2.2697	5.8407	4.4717	7.2499
0.1860	2.6106	0.8663	3.2878	2.2958	4.5960	4.5497	3.8297
0.1872	0.8986	0.8665	2.5083	2.3071	3.2522	4.5565	6.3410
0.1995	1.3440	0.8735	2.6362	2.3105	5.6482	4.6005	16.6148
0.2563	1.3911	0.8801	3.6306	2.3434	9.4699	4.6071	7.8126
0.2846	2.0412	0.8845	5.1947	2.3873	4.3348	4.7358	3.7244
0.3126	1.5188	0.9333	2.0952	2.4731	6.6499	4.7994	6.5136
0.3334	0.0000	0.9431	3.0799	2.4767	5.7450	4.8478	5.0103
0.3374	1.2937	0.9559	3.3119	2.5148	3.2558	4.9112	12.0885
0.3435	1.6993	0.9942	5.7759	2.5737	5.4331	5.0054	7.9832
0.3710	2.7574	1.0000	0.7500	2.5845	4.8119	5.1690	7.3258
0.3805	0.8898	1.0000	0.0497	2.5897	3.7948	5.1761	10.2428
0.3889	0.8572	1.0691	1.9937	2.6169	5.3518	5.3629	11.6814
0.3894	1.8671	1.0876	2.3614	2.6176	4.2769	5.6096	7.8623
0.4000	0.6750	1.0964	3.5785	2.6176	4.2472	5.6107	7.5192
0.4127	1.7768	1.1374	2.9386	2.7574	5.0218	5.6261	8.0917
0.4489	1.4567	1.1490	1.8510	2.7607	2.1175	5.8070	8.5122
0.4708	3.5397	1.1500	2.9739	2.8217	3.2629	6.1159	13.9024
0.4991	2.6353	1.1692	1.2862	2.8421	2.6389	6.7663	12.8395
0.5000	0.0000	1.2381	1.6987	2.8468	9.8046	6.9818	6.6899
0.5476	1.5653	1.2568	4.1001	2.8982	3.4162	6.9968	15.0863
0.5551	1.7282	1.2571	2.0097	2.9059	4.1463	7.6552	10.7643
0.5701	2.8392	1.3498	2.1599	2.9938	8.9665	7.8442	15.6844
0.5744	2.6857	1.3741	6.8814	3.0893	4.2798	7.8893	10.9467
0.6182	0.4235	1.4365	5.2842	3.1953	9.6635	7.9251	6.8570
0.6182	0.4235	1.4660	4.3234	3.3971	1.4596	8.5472	13.0823
0.6378	1.3918	1.4698	3.0077	3.4404	6.2161	9.5597	10.5194
0.6500	4.4973	1.4866	3.6210	3.4517	6.2397	13.6703	18.3448
0.6517	3.6513	1.5408	4.0441	3.6323	3.6982	13.7195	25.0399
0.6781	4.4256	1.5613	4.9363	3.6922	8.1021	14.6661	24.7414
0.6786	1.1843	1.5719	4.5029	3.7623	4.2254	19.6395	46.9552
0.7224	4.2417	1.6439	1.6936				

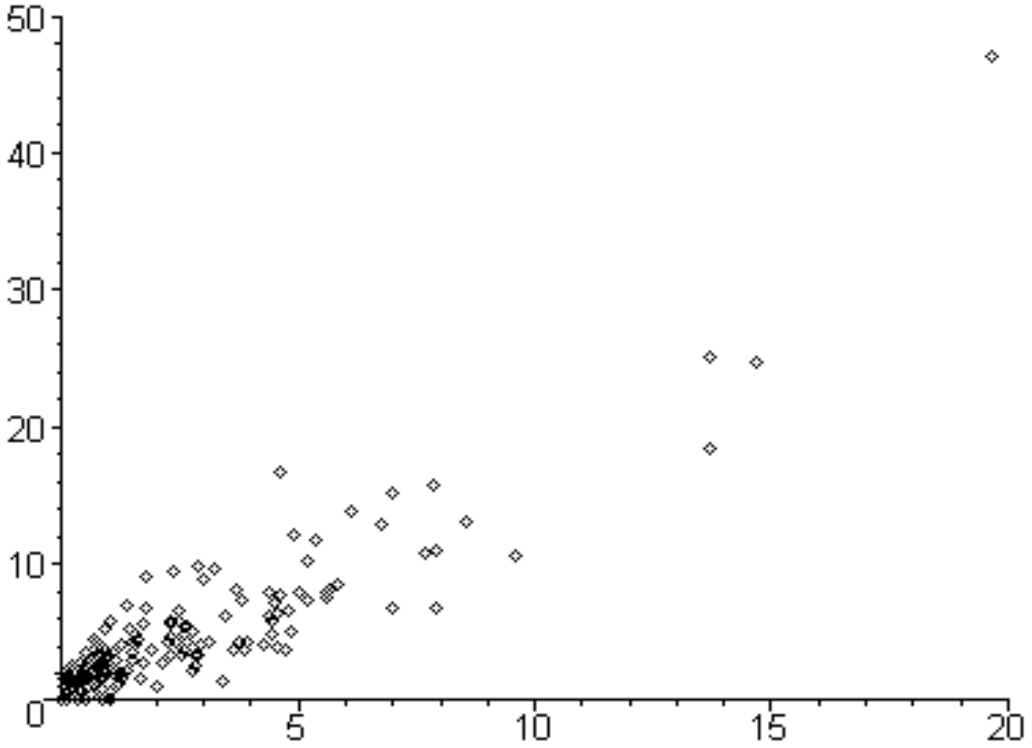


Figure 2. The <I,S> points of meaning diversification

However, there is no unique picture and some of the distributions obtained are quite complex. In order to simplify and unify the procedure we shall apply here Popescu´s approach (cf. Popescu, Altmann, Köhler 2008). It is especially appropriate in cases where rankings result from separate development in strata, e.g. in different word classes, and the strata are put together. Here we model the rank-frequency data using a continuous function. The approach is elementary and in the first step one simply assumes a constant relative rate of change (decrease) of frequencies. Since frequencies are taken into account only if $f_r \geq 1$, this value is set as limit. Thus one obtains

$$\frac{df_r}{f_r - 1} = -b \tag{4}$$

Solving (4) yields

$$f_r = 1 + ae^{-br}. \tag{5}$$

This approach is a special case of the general theory (Wimmer, Altmann 2005) hence at least this first step is systematized. The continuation consisting of simple adding of further exponential components $f_r = 1 + ae^{-br} + ce^{-dr} + \ldots$ requires non-homogeneous differential equations of higher order and deviates from the general theory. However, it has a great advantage in word frequency studies: if the exponents of two components are equal, one of them can be eliminated. In this way the formula shows not only the extent of diversification but also that of stratification. Popescu et al. (2008) used the procedure for rank-frequency

distributions of words in texts. Formula (5) is analogous to Zipf´s law which, however, contains a power function.

Let us illustrate the problem using the semantic diversification of *Fly* as shown in Table 5.

Table 5
Fitting (5) to *Fly*

r	f_r	One component	Two components
1	33	32.55	33.00
2	9	11.30	9.04
3	6	4.37	6.19
4	5	2.10	4.35
5	3	1.36	3.17
6	2	1.12	2.40
7	2	1.04	1.90
8	2	1.01	1.58
9	1	1.00	1.38
10	1	1.00	1.24
11	1	1.00	1.16
12	1	1.00	1.10
13	1	1.00	1.07
14	1	1.00	1.04
15	1	1.00	1.03
16	1	1.00	1.02
17	1	1.00	1.01
18	1	1.00	1.01
19	1	1.00	1.00
20	1	1.00	1.00
		a = 96.5724 b = 1.1188 D = 0.97	a = 19.2642 b = 0.4370 c = 420964.899 d = 9.9771 D ≈ 1

Though the fitting signalizes that *Fly* has two components, to capture its course one component is sufficient. And this “mono-componentiality”. i.e. homogeneous diversification in all word classes is characteristic for all words in the sample. While preliminarily it is impossible to find a unique discrete distribution capturing the diversification of all English words, function (5) is adequate in all cases, as shown in Table 6 containing the given word, the parameters *a* and *b* and the determination coefficient *D*.

Table 6
Fitting of (5) to semantic diversification of English words

Word	a	b	D	Word	a	b	D
Animal	498472.593	8.9297	1	Louse	8534.5925	9.0519	1
Ash	8.535	9.0519	1	Man	3981.7381	1.0299	0.99
Back	131.4681	0.4156	0.98	Meat	26245.3705	9.0766	1

Bad	1151.2902	3.1367	1	Moon	254016.475	9.0779	1
Bark	21992.7741	8.8999	1	Mother	9802.9907	4.5953	1
Belly	50.9433	1.9844	1	Mountain	131013.174	9.0105	1
Bird	233.790.3940	8.961	1	Mouth	211.5569	1.4831	1
Bite	122.976	2.414	1	Name	10273.7638	2.6908	1
Black	894.8582	2.7896	0.99	Near	103.6163	0.8756	0.99
Blood	19434.3109	3.4196	1	Neck	260677.389	8.9745	1
Blow	61.3284	0.9664	0.96	New	2488.9156	0.9457	0.99
Bone	44.2503	1.5908	1	Night	2598.5192	1.2615	1
Breast	26.896	1.6812	1	Nose	424.472	2.6835	1
Breathe	197842.642	9.0172	1	Old	1081.1364	0.6566	0.9
Burn	16.4593	0.3563	0.91	Play	78.1554	0.2539	0.97
Child	2386.9828	1.3372	0.99	Pull	204.7085	1.5657	0.97
Cloud	57.291	0.8651	0.93	Push	416.6638	2.0265	0.99
Cold	118.9857	1.1193	0.99	Rain	56.7698	0.8444	0.98
Come	474.3732	0.4594	0.96	Red	143.3153	1.2435	0.93
Correct	25.9069	0.6598	0.95	Right	3913.6907	1.8011	0.98
Count	61.0688	1.0392	0.95	River	-	-	-
Cut	19023.3662	2.4325	1	Road	8837.9997	4.5435	1
Day	1323.7953	0.6888	0.98	Root	42.0852	1.4419	0.99
Die	3986.1812	3.3418	1	Rope	32563.4575	9.0047	1
Dig	22.3923	1.0051	0.96	Rotten	1	0	1
Dirty	122.976	2.414	1	Round	26.4228	0.8002	0.99
Dog	1682.9982	3.7148	1	Rub	138579.338	8.949	1
Drink	60.4861	0.641	0.99	Salt	163.9707	1.8802	1
Dry	35.7825	0.6013	0.94	Sand	82.9642	2.221	1
Dull	7.164	0.475	0.91	Say	9945.5387	1.3426	1
Dust	135.6992	1.191	1	Scratch	18.3661	0.7404	0.86
Ear	163.9656	1.5485	0.98	Sea	688.4826	2.9236	1
Earth	126.1833	0.8285	0.96	See	1620.5453	0.9907	0.96
Eat	1622.4235	1.2208	1	Seed	37.1791	1.208	0.99
Egg	138597.338	8.9494	1	Sew	31045.6339	8.9569	1
Eye	6554.6188	3.2158	1	Sharp	10.5661	0.2839	0.94
Fall	70.9576	0.5119	0.97	Sing	103.6247	0.7684	0.87
Far	104.2227	0.5212	0.98	Sit	536.0568	1.3949	1
Fat	398.0246	2.9421	1	Skin	28.8576	1.0861	0.93
Father	1636.0032	3.1374	1	Sky	385171.853	8.9902	1
Fear	177.0005	0.9169	0.98	Sleep	185.2828	1.1673	0.98
Feather	39308.6176	8.9698	1	Smell	21.4613	0.4141	0.94
Fight	388.3263	0.4193	0.96	Smoke	395.6359	1.7762	1
Fingernail	-	-	-	Smooth	22.252	0.7645	0.97
Fire	2590.2801	1.446	0.97	Snake	184.4363	2.2728	1
Fish	76.3803	1.7721	1	Snow	21.5563	0.4775	0.79
Float	47.7162	1.3054	0.99	Spit	73827.4163	8.9069	1
Flow	26.6174	0.3884	0.97	Split	6.1677	0.3971	0.97
Flower	141.58	1.5494	1	Squeeze	18.6917	0.706	0.98

Fly	96.5724	1.1188	0.98	Stab	1	0	1
Fog	91.789	1.687	1	Stand	406.9846	0.9043	0.99
Foot	1790.6843	0.8548	0.97	Star	14.1537	0.7504	0.92
Forest	1226.9951	3.557	1	Stick	9.0738	0.3214	0.94
Freeze	12.195	0.6599	0.96	Stone	686.7294	0.741	1
Fruit	65.846	1.9918	0.99	Straight	19.7573	0.3971	0.98
Full	87.2529	0.6913	0.94	Suck	10.7633	1.2702	0.99
Give	264.166	0.3229	0.97	Sun	198.385	1.4583	1
Good	1514.6762	2.0821	0.99	Swell	6.2325	0.36	0.91
Grass	320486.465	8.9887	1	Swim	86915.2607	8.9748	1
Green	134.4803	1.6813	1	Tail	26.8961	1.6813	1
Guts	2.2438	0.6909	0.77	Think	535.717	0.6106	0.96
Hair	480029.837	9.0212	1	Throw	255.748	1.5984	0.95
Hand	1894.3731	2.1761	1	Tie	17.6175	0.4896	0.92
Head	1242.6036	1.795	0.99	Tongue	54.5538	1.435	1
Hear	1275.3905	1.5379	1	Tooth	1370.9978	3.6124	1
Heart	86.4807	0.721	0.99	Tree	839597.631	8.9772	1
Hit	571.3084	0.2964	0.96	Turn	1326.7787	0.4997	0.96
Hold	1420.8213	0.3453	0.96	Vomit	1	0	1
Horn	21.527	1.2702	0.99	Walk	31036.0977	3.3481	1
Hunt	10.7635	1.2702	0.99	Warm	146.793	1.4941	1
Husband	516642.638	8.921	1	Wash	18.2339	0.4	0.98
Ice	274287.39	9.1207	1	Water	2652.4829	1.2706	1
Kill	5205.9977	3.9326	1	Wet	167.1774	2.0277	1
Knee	1253.981	3.2221	1	White	235.4116	1.3097	0.99
Know	1700.1555	1.0738	0.97	Wife	-	-	-
Lake	18305.8794	9.1218	1	Wind	365.2692	2.5686	1
Laugh	300.9258	1.5457	1	Wing	10.8622	0.3861	0.91
Leaf	147259.259	8.9555	1	Wipe	131017.174	9.0105	1
Left	206.6777	0.4261	0.94	Woman	2553.4173	1.6727	1
Leg	900.3727	2.4988	1	Worm	5.5921	1.0041	0.95
Lie	166.8116	0.6204	0.97	Year	26257.0274	3.4569	1
Live	288.9317	0.8016	0.99	Yellow	93.8062	1.3208	1
Liver	82535.9427	9.0184	1				

In Table 6, the determination coefficient D has been rounded to two decimal places, hence we frequently got 1.00. In general, the fitting is better than $D = 0.9$, i.e. the result is excellent. In three cases which did not display diversification (*Fingernail, River, Wife*), the formula could not be applied. In the three cases displaying uniform distribution (*Rotten, Stab, Vomit*), parameter b is set at 0 and parameter a at 1 because there is maximal diversification (no concentration); all senses are equally frequent. (the same result can be achieved with $a = 0$). In three cases the determination coefficient is $D \in \langle 0.79, 0.86 \rangle$ (*Guts, Scratch, Snow*) but in general all results are satisfactory.

Parameter a is an equilibrating quantity; parameter b shows an aspect of diversification. In cases of maximal concentration, where all but the first $f_r = 1$ and the first class represents the strongly emphasized primary meaning, parameter $b > 8.00$. That means, the

function decreases very abruptly from some frequencies greater than 1 to 1 representing all other senses, e.g. *Ash*: 2,1,1,1; *Animal*: 67,1,1; *Bark*: 4,1,1,1,1,1,1,1,1; *Breathe*: 25,1,1,1,1,1,1,1,1, etc. Thus a large b means strong concentration, $b = 0$ no concentration. Comparing b with R and eliminating cases in which $b = 0$ and $b > 8$, b and R are almost linearly correlated. The number of cases, as compared with the bulk of language, is of course too small to allow a better interpretation of the parameter b.

Thus instead of searching for a unique discrete distribution representing the rank-frequency of individual meanings one can use the very simple Popescu function remembering decay with an infinite tail. Theoretically, we let the tail be infinitely long but, as mentioned above, the hearers control its length. Nevertheless, it is not possible to predict its length.

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.

Hennern, A. (1991). Zur semantischen Diversifikation von „in“ im Englischen. In: Rothe (1991), 116-126.

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

Levickij, V.V. (2005). Polysemie. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 458-464*. Berlin/New York: de Gruyter.

Mačutek, J. (2007). Pairs of corresponding discrete and continuous distributions: Mathematics behind, algorithms and generalizations. In: Köhler, R., Grzybek, P (eds.), *Exact Methods in the Study of Language and Text: 407-414.* Berlin/New York: de Gruyter.

Mačutek, J. (2008). On the distribution of graphemic representations. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties and characters of writing systems: 75-78.* Berlin/New York: Mouton de Gruyter.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modeling in quantitative linguistics. *Journal of Quantitative Linguistics 14, 2007, 81-94.*

Popescu, I.-I., Altmann, G., Köhler, R. (2008). Zipf´s law – another view (s*ubmitted*).

Rothe, U. (ed.) (1991). *Diversification process in language: grammar*. Hagen: Rottmann.

Shenton, I.R., Skees, P. (1970). Some statistical aspects of amounts and duration of rainfall. In: Patil, G.P. (ed.), *Random counts in scientific work, Vol 3: 73-94*. University Park: The Pennsylvania State University Press.

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

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

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

Zipf, G.K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology 33, 251-256.*

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

Glottometrics 17, 2008, 79-86

Arc length and meaning diversification in English

Fengxiang Fan, Dalian[1]
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. We try to characterize meaning diversification using the relationship of some properties of the empirical diversification distribution and defining a coefficient *c* whose mean holds true for all examined words and which allows us to compute the theoretical arc length.

Keywords: Diversification, English, arc length

In a previous article we presented some properties of semantic diversification of 165 English words, i.e. the rank-frequency distribution of individual senses of words. The data were taken randomly from a dictionary and their frequencies from WodNet[2] (cf. Fan, Altmann 2008). First we captured the distribution by a number of discrete mass functions, but none of them held in all cases. Hence we approximated the distributions using Popescu´s alternative for Zipf´s law (Popescu et al. 2008) and obtained a significant result in all cases. We used the formula

$$(1) \qquad f_r = 1 + ae^{-br}$$

where parameter *b* shows the extent of *semantic concentration* and *a* is an equilibrating constant. The semantic concentration could be measured also using the usual Repeat rate or entropy.

The approximation of seemingly discrete data by a continuous function is no blemish because (1) in nature there are no discrete or continuous phenomena, only our concepts are so; (2) discretization in semantics is always a kind of very artificial but practically very simplifying measure allowing us to capture meaning phenomena. Hence returning to a continuous function is the same kind of dodge as using discrete rank-orders in semantics. Besides, the use of the determination coefficient enables us to avoid the problems associated with the use of the chi-square criterion for small numbers.

The semantic concentration of meaning is, so to say, the opposite of diversification.

In this contribution we want to observe the arc length built by the ranked frequencies of senses, especially its evolution in the course of diversification. Arc length has already been used for the characterization of word frequencies (cf. Popescu, Mačutek, Altmann 2008). For continuous functions it is defined as (*S* = number of different senses)

$$(2) \qquad L = \int_1^S \sqrt{1 + y'^2}\, dx,$$

yielding after integration a voluminous but simple result. For a sequence of points around *y* we get

[1] Address correspondence to: fanfengxiang@yahoo.com
[2] (http://www.cogsci.princeton.edu/~wn/

$$(3) \qquad L = \sum_{r=1}^{S-1} \{[f(r) - f(r+1)]^2 + 1\}^{1/2}$$

and (3) yields slightly larger numbers than (2) caused by the deviations of points from *y*. Here we shall use (3), which can easily be computed from our data. If $S = 1$ (as is the case of the words *Fingernail*, *River*, and *Wife* in Table 1 bellow), there is no arc and we define $L = f_1$ without distorting the results. Consider for example the ordered frequencies of the word Belly having S = 6 meanings given as 8, 2, 1, 1, 1, 1 yielding

$$L = [(8-2)^2+1]^{1/2} + [(2-1)^2+1]^{1/2} + 1 + 1 + 1 = 10.50.$$

We conjecture that the development of diversification has a very regular course. It can begin in two forms: (i) either the main (the unique) meaning is originally very concentrated and occurs in all texts (environments), while secondary senses begin to develop when it acquires high frequency (e.g. 100, 1, 1). (ii) The word already arises with diffuse meaning having different senses (e.g. with frequencies 1, 1, 1, 1) and only later on one of them achieves the central role. In our data all possible development stages can be observed.

But once the diversification is started, a kind of self-organization comes into existence and compels the arc to enter an attractor bestowing it with a prescribed length. It is a kind of constant relationship – many of which are known from physics, mutatis mutandis. In word frequency research it could be shown (Popescu, Mačutek, Altmann 2008) that arc length (L), vocabulary size (V), the greatest frequency (f_1) and the h-point abide by the relationship

$$(4) \qquad L - V - f_1 + 1.3h = 0.$$

The multiplicative constant (1.3) with h is more exactly $c = 1.297 \pm 0.0888$ and we suppose that it is given by some boundary conditions concealed in the level but not in the language, and not in the genre either, because in 100 texts of different genres and from different times in 20 languages the same result has been obtained. We suppose that this relationship holds for different levels of language where ranking may be established and the only difference is in the constant c. Here we shall show this relationship applied to meaning diversification of 165 English words. Table 1, lists165 randomly sampled English words whose meaning diversification has been obtained from WordNet giving information also about the frequency of the given sense. To each word the arc length (L), the number of meanings (senses, S), the greatest frequency (f_1) and the h-point computed by standard method (cf. Popescu et al. 2008)[3] are given. For the sake of simplicity we compute the arc length L_o using formula (3) and the expected L_t using

$$(5) \qquad L_t = S + f_1 - ch$$

where c is computed as the mean of all results in the seventh column. As can be seen, for individual words c oscillates slightly – perhaps according to the present state of diversification – but the mean of all c-values is 1.4699 with standard deviation 0.1681. Rounding to 1.47 we obtain the computed values $L_t = S + f_1 - 1.47h$ in the last column of Table 1. The det-

[3] If an $f_r = r$, then $h = r$. If not, one takes two neighbouring ranks.$r_2 = r_1 + 1$ such that $r_1 < f_1$ and $r_2 > f_2$, Then by interpolation $h = \dfrac{f_1 - (f_2 - f_1)r_1}{1 - f_2 + f_1}$

ermination coefficient is $R = 1$. Thus, the relationship is very stable. In Figure 1 the graphical comparison shows the perfect agreement of observed and hypothesized L (formula 5).

Table 1
Computing arc length for meaning diversification of 165 English words.

WORD	$N = \Sigma f_i$	L_o	f_1	S	h	$c = \frac{S + f_1 - L}{h}$	$L_t = f_1 + S - 1.47*h$
Animal	69	67.01	67	3	1.98	1.5113	67.09
Ash	5	3.41	2	4	1.50	1.7239	3.80
Back	302	109.19	92	28	8.66	1.2484	107.27
Bad	72	63.84	51	17	3.00	1.3871	63.59
Bark	12	10.16	4	9	1.75	1.6216	10.43
Belly	14	10.50	8	6	2.00	1.7515	11.06
Bird	36	34.02	31	6	1.92	1.5538	34.18
Bite	25	21.46	12	13	2.00	1.7680	22.06
Black	91	74.25	56	23	4.00	1.1869	73.12
Blood	677	637.35	637	6	3.91	1.4442	637.25
Blow	72	47.92	25	29	4.33	1.4040	47.63
Bone	17	12.31	10	6	2.33	1.5849	12.57
Breast	12	8.54	6	6	2.00	1.7313	9.06
Breathe	33	31.02	25	9	1.94	1.5357	31.15
Burn	55	26.09	11	20	4.25	1.1554	24.75
Child	823	622.09	625	4	3.86	1.7910	623.33
Cloud	51	33.10	24	13	2.87	1.3604	32.78
Cold	75	51.20	40	16	4.20	1.1419	49.83
Come	792	286.15	275	22	7.67	1.4141	285.73
Correct	40	22.05	15	12	4.25	1.1656	20.75
Count	52	28.44	23	11	4.00	1.3910	28.12
Cut	2138	1728.05	1672	71	10.74	1.3915	1727.21
Day	1314	648.65	648	10	7.25	1.2899	647.34
Die	160	152.10	142	14	2.67	1.4597	152.08
Dig	23	16.25	9	11	2.60	1.4438	16.18
Dirty	25	21.46	12	13	2.00	1.7680	22.06
Dog	50	46.43	42	8	2.00	1.7866	47.06
Drink	74	35.82	32	10	4.00	1.5439	36.12
Dry	59	34.53	20	19	3.73	1.1994	33.52
Dull	30	20.58	5	19	2.75	1.2449	19.96
Dust	64	44.30	42	7	3.00	1.5679	44.59
Ear	51	36.53	36	5	3.50	1.2770	35.86

Earth	105	61.28	57	9	3.94	1.1982	60.21
Eat	680	478.22	479	6	5.17	1.3108	477.40
Egg	23	21.03	19	5	1.92	1.5480	21.18
Eye	291	264.50	264	6	4.33	1.2704	263.63
Fall	169	81.90	46	44	5.33	1.5204	82.16
Far	155	64.79	62	10	5.00	1.4415	64.65
Fat	34	28.44	22	10	2.00	1.7804	29.06
Father	86	76.66	72	9	2.66	1.6325	77.09
Fear	127	75.29	73	8	4.43	1.2882	74.49
Feather	12	10.10	6	7	1.83	1.5852	10.31
Fight	729	263.86	268	9	7.50	1.7523	265.98
Fingernail	1	1.00	1	1	1.00	1.0000	0.53
Fire	1017	620.87	616	17	9.25	1.3115	619.40
Fish	22	15.87	14	6	2.50	1.6505	16.33
Float	33	24.70	14	15	3.00	1.4333	24.59
Flow	66	25.12	18	14	5.00	1.3767	24.65
Flower	41	31.09	31	4	2.75	1.4208	30.96
Fly	74	46.66	33	20	4.34	1.4604	46.62
Fog	25	17.69	18	4	2.67	1.6157	18.08
Foot	1282	745.23	740	14	6.00	1.4617	745.18
Forest	39	35.43	36	3	2.00	1.7855	36.06
Freeze	26	16.71	7	14	3.00	1.4306	16.59
Fruit	16	11.48	10	5	2.00	1.7618	12.06
Full	94	49.02	42	13	3.80	1.5733	49.41
Give	805	206.70	181	42	13.33	1.2230	203.40
Good	303	207.53	190	27	7.50	1.2628	205.98
Grass	50	48.01	41	10	1.96	1.5242	48.12
Green	44	36.12	26	14	2.67	1.4517	36.08
Guts	8	5.41	2	6	2.00	1.2929	5.06
Hair	64	62.01	59	6	1.93	1.5499	62.16
Hand	265	225.68	216	16	4.00	1.5789	226.12
Head	337	241.75	208	42	6.00	1.3754	241.18
Hear	356	274.21	275	5	4.50	1.2874	273.39
Heart	88	45.76	42	10	4.25	1.4682	45.75
Hit	1627	449.55	440	24	12.00	1.2039	446.36
Hold	3906	1154.76	1134	45	19.00	1.2757	1151.07
Horn	19	14.36	7	11	2.33	1.5626	14.57
Hunt	19	15.65	4	15	2.00	1.6749	16.06
Husband	71	69.01	70	2	1.83	1.6339	69.31
Ice	40	38.02	31	10	1.95	1.5299	38.13

Kill	121	116.24	103	17	2.33	1.6133	116.57
Knee	55	50.25	51	3	2.33	1.6110	50.57
Know	968	597.17	593	12	6.40	1.2228	595.59
Lake	5	3.24	3	3	1.67	1.6550	3.55
Laugh	83	65.04	65	4	2.87	1.3786	64.78
Leaf	25	23.03	20	6	1.91	1.5569	23.19
Left	485	161.43	151	24	11.43	1.1871	158.20
Leg	90	79.52	75	9	2.84	1.5773	79.83
Lie	208	91.24	89	10	6.50	1.1938	89.45
Live	264	144.63	133	19	6.00	1.2276	143.18
Liver	15	13.05	11	5	1.91	1.5446	13.19
Louse	5	3.41	2	4	1.50	1.7239	3.80
Man	2283	1441.57	1437	13	6.00	1.4050	1441.18
Meat	6	4.16	4	3	1.75	1.6216	4.43
Moon	38	36.02	30	9	1.91	1.5617	36.19
Mother	107	103.42	100	7	2.00	1.7903	104.06
Mountain	18	16.03	17	2	1.97	1.5076	16.10
Mouth	74	54.90	49	11	3.00	1.7009	55.59
Name	847	703.70	698	15	6.00	1.5506	704.18
Near	80	48.13	44	9	3.73	1.3051	47.52
Neck	38	36.02	34	5	1.91	1.5627	36.19
New	1648	982.99	980	12	6.00	1.5009	983.18
Night	1041	735.80	736	8	6.29	1.3038	734.75
Nose	45	40.25	30	14	2.33	1.6075	40.57
Old	1066	516.96	515	9	4.60	1.5294	517.24
Play	331	109.34	70	52	8.67	1.4600	109.26
Pull	90	62.24	44	24	5.00	1.1519	60.65
Push	88	65.31	56	15	4.00	1.4223	65.12
Rain	44	24.23	25	4	3.40	1.4038	24.00
Red	79	45.20	43	8	5.40	1.0736	43.06
Right	1032	670.49	649	35	11.75	1.1496	666.73
River	55	55.00	55	1	1.00	1.0000	54.53
Road	99	95.42	95	4	2.00	1.7902	96.06
Root	29	21.89	11	15	2.50	1.6437	22.33
Rope	8	6.12	5	4	1.80	1.5983	6.35
Rotten	3	2.00	1	3	1.00	2.0000	2.53
Round	48	34.06	13	26	3.00	1.6461	34.59
Rub	23	21.03	19	5	1.91	1.5561	21.19
Salt	39	32.15	26	10	2.60	1.4820	32.18
Sand	14	10.48	10	4	2.00	1.7618	11.06

Say	3547	2593.96	2593	12	8.20	1.3462	2592.95
Scratch	29	19.49	9	14	2.75	1.2781	18.96
Sea	43	38.25	38	4	2.33	1.6093	38.57
See	1227	626.97	617	25	12.00	1.2525	624.36
Seed	28	21.19	12	13	2.60	1.4638	21.18
Sew	6	4.12	5	2	1.80	1.6000	4.35
Sharp	45	19.15	9	15	4.50	1.0771	17.39
Sing	86	46.30	46	5	3.00	1.5676	46.59
Sit	187	136.25	134	8	4.33	1.3282	135.63
Skin	28	17.89	11	11	3.00	1.3698	17.59
Sky	50	48.01	49	2	1.83	1.6339	48.31
Sleep	85	60.04	58	6	2.91	1.3619	59.72
Smell	46	15.99	14	8	4.50	1.3348	15.39
Smoke	91	73.23	68	10	3.25	1.4666	73.22
Smooth	30	18.48	11	12	3.00	1.5059	18.59
Snake	29	24.27	20	8	2.33	1.6028	24.57
Snow	37	14.71	13	6	3.70	1.1606	13.56
Spit	18	16.05	11	8	1.91	1.5446	16.19
Split	31	19.66	5	19	3.00	1.4477	19.59
Squeeze	34	22.52	10	17	3.25	1.3780	22.22
Stab	6	5.00	1	6	1.00	2.0000	5.53
Stand	330	183.89	169	24	6.60	1.3800	183.30
Star	25	16.34	8	12	3.00	1.2216	15.59
Stick	48	28.40	7	26	4.24	1.0837	26.77
Stone	629	338.61	330	16	5.00	1.4776	338.65
Straight	59	28.69	14	21	5.00	1.2619	27.65
Suck	10	6.65	4	6	2.00	1.6749	7.06
Sun	65	50.06	47	7	2.84	1.3886	49.83
Swell	24	11.66	5	11	3.50	1.2409	10.86
Swim	14	12.05	12	3	1.91	1.5469	12.19
Tail	17	13.54	6	11	2.00	1.7313	14.06
Think	602	283.38	277	14	4.80	1.5869	283.94
Throw	110	66.14	53	20	5.50	1.2472	64.92
Tie	47	25.72	13	18	4.00	1.3202	25.12
Tongue	27	19.70	14	10	2.67	1.6104	20.08
Tooth	43	39.43	38	5	2.00	1.7860	40.06
Tree	113	111.00	107	7	1.91	1.5682	111.19
Turn	2091	765.49	744	38	12.25	1.3474	763.99
Vomit	4	3.00	1	4	1.00	2.0000	3.53
Walk	1208	1099.55	1092	17	6.75	1.4001	1099.08

Warm	57	42.35	34	13	3.00	1.5486	42.59
Wash	56	27.76	13	21	4.50	1.3858	27.39
Water	1026	747.84	744	10	3.98	1.5470	748.15
Wet	34	28.19	23	9	2.50	1.5246	28.33
White	117	84.33	65	25	4.33	1.3092	83.63
Wife	120	120.00	120	1	1.00	1.0000	119.53
Wind	47	39.85	29	15	2.50	1.6609	40.33
Wing	31	13.23	8	10	4.25	1.1231	11.75
Wipe	18	16.03	17	2	1.93	1.5389	16.16
Woman	587	475.24	480	4	4.00	2.1892	478.12
Worm	8	4.83	3	5	2.00	1.5858	5.06
Year	865	831.15	832	4	3.40	1.4275	831.00
Yellow	42	29.52	26	8	2.86	1.5648	29.80

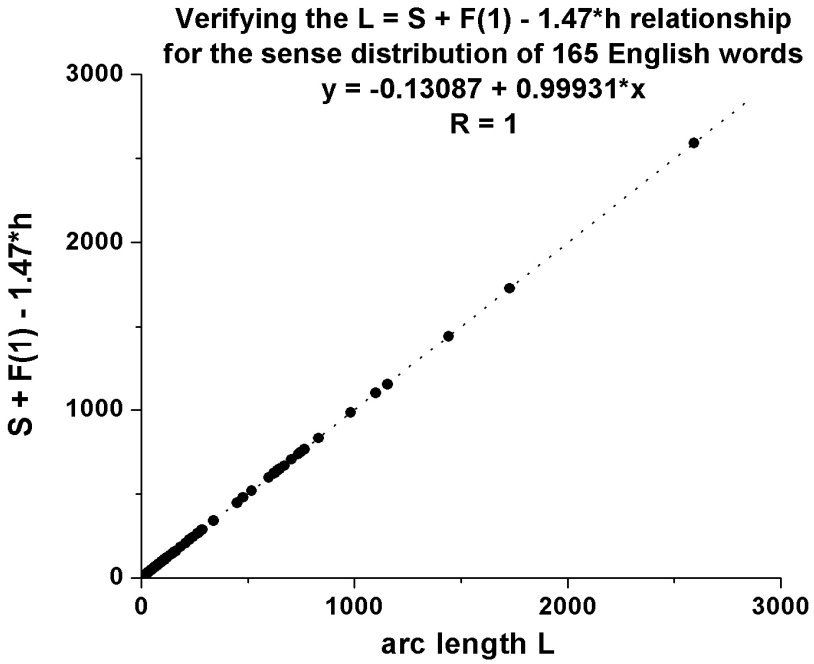


Figure 1. Arc length and other quantities of meaning diversification

The results show that there is a kind of "prescribed" development of meaning diversification. Nevertheless, a number of problems remain open and must be scrutinized in future research: (1) Does this relationship hold only for English or can be found in other languages, too? (2) The parameter *c* is represented by its mean, however, its value differs for word-frequency distributions and meaning diversifications. In diversification it is slightly greater. The question is, why? Can we conjecture that the more concrete the linguistic level, the smaller is *c*? That is, is there an increase in *c* beginning from phonemes, to syllables, words, morphemes, morpheme classes (e.g. prepositions), meanings, meaning classes (e.g. colours, grammatical categories)? How can *c* be interpreted? Is it associated with redundancy or other requirements that must be fulfilled by language (cf. Köhler 2005)?

The fact that the *h*-point plays here an eminent role is a further evidence of its reasonability. It is to be expected that in the future further relations will be discovered.

References

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 69-81.*

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-775.* Berlin/New York: de Gruyter

Popescu, I.-I., Altmann, G., Köhler, R. (2008). Zipf´s law – another view (submitted).

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-44.*

Popescu, I.-I. et al. (2008). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Glottometrics 17, 2008, 87-93

Sinismen im Deutschen und Englischen

Karl-Heinz Best, Göttingen[1]

Abstract. The paper deals with two topics: The process of borrowing of Chinese words in German and English. It can be shown that the borrowings in English abide by the Piotrowski law (logistic law).

Keywords: Borrowings, Chinese, English, German, Piotrowski law

Entlehnungen als Gegenstand der Quantitativen Linguistik

In diesem Beitrag geht es um die Entlehnungen aus dem Chinesischen (Sinismen) ins Deutsche und Englische. Die leitende Hypothese dieser Untersuchung formulieren Strauss, Fan & Altmann (2008: 36): „The number of loan words increases in every language according to Piotrowski law." Das Piotrowski-Gesetz entspricht dem logistischen Modell in der Form

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

und hat sich in einer Vielzahl von Untersuchungen bereits bewährt (Best 2001, Körner 2004 und viele andere; zur Begründung des Modells siehe Altmann 1983.).

Als Sinismen werden alle die Wörter betrachtet, die entweder letztlich aus dem Chinesischen als Herkunftssprache stammen oder über das Chinesische als Vermittlersprache gekommen sind. Zwischen Fremd- und Lehnwörtern wird nicht unterschieden; einzelne Lehnübersetzungen wurden berücksichtigt.

In bisherigen Untersuchungen zu den Entlehnungen ins Deutsche spielen die Sinismen nur eine sehr geringe Rolle. Ein deutlich anderes Bild gewinnt man, wenn man Cannons Untersuchung zum Englischen betrachtet: Er kommt auf insgesamt 1191 Sinismen (Cannon 1988: 4). Man muss nun nicht erwarten, dass sich im Deutschen ähnlich viele Entlehnungen finden; die wenigen Einzelfälle, die man in der Regel kennt (Dschunke, Kotau, Taifun) sind aber bei Weitem nicht alle.

Zwei Themen werden daher behandelt: Es sollen erstens möglichst viele im Deutschen gebräuchliche Sinismen erfasst und aufgelistet werden. Die ursprüngliche Absicht war, an die gewonnenen Daten das Modell (1) anzupassen. Leider hat sich herausgestellt, dass nur sehr wenige Sinismen datiert werden können. Daher wird hier auf eine Modellierung des Übernahmeprozesses vorerst verzichtet.

Beim zweiten Thema geht es um die Sinismen im Englischen. Sie müssen nicht gelistet werden, da Cannon (1988: 25-31) sie schon in einem Anhang aufführt. Stattdessen wird aus seinen chronologischen Beschreibungen (Cannon 1988: 4f.) in einer Tabelle die Dynamik der Entlehnungen erfasst und zugleich die Anpassung von Modell (1) an die Daten dargestellt. Man darf bis zum Beweis des Gegenteils vermuten, dass der so dargestellte Trend im Deutschen ähnlich verläuft, wenn auch auf niedrigerem Niveau.

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

Entlehnungen aus dem Chinesischen ins Deutsche

Der folgende Abschnitt ist den Wörtern gewidmet, die ins Deutsche entlehnt wurden und auf das Chinesische zurückgeführt werden können. Entlehnungen im Deutschen stammen aus über 30 Sprachen; wahrscheinlich sind es wesentlich mehr, da die entsprechenden Angaben in (Best 2005) sich auf Auswertungen etymologischer Wörterbücher sowie eines nicht mehr ganz aktuellen Fremdwörterbuchs stützt. Sinismen spielen dabei eine ganz untergeordnete Rolle; die Angaben schwanken zwischen 1 und 3 Wörtern chinesischen Ursprungs (Best 2001, Körner 2004). Tatsächlich lassen sich in neueren Wörterbüchern über 100 solcher Einträge nachweisen; nähme man alle Wortbildungen hinzu, in denen Konstituenten aus dem Chinesischen stecken, wären es noch deutlich mehr.

Die Untersuchung beruht auf einer Auswertung der etymologischen Wörterbücher von *Duden. Herkunftswörterbuch* (2001), Kluge (2002), und Pfeifer (1995) sowie als ergiebigere Quellen *Duden. Das große Wörterbuch* (1999) und *Duden. Das große Fremdwörterbuch* (2007).

Tabelle 1 enthält Wörter, in denen mindestens ein Wortteil aus dem Chinesischen stammt. Außerdem sind insgesamt drei Lehnübersetzungen aufgenommen, die auf chinesische Bezeichnungen zurückgehen. Reine Eigennamen sind nicht berücksichtigt, wohl aber Namen, die als Gattungsbezeichnungen verwendet werden oder als Wortteil in allgemein gebräuchliche Wörter eingegangen sind. So fehlt z.B. „Hsinhua“, die Bezeichnung für die chinesische Nachrichtenagentur, nicht aber „Nanking“, das nicht nur die betreffende Stadt, sondern eben auch ein Gewebe bezeichnet. Entsprechend fehlt „Mao“, nicht aber „Maoismus“.

Die Darstellung ist unvollständig, da zu etlichen der aufgenommenen Wörter weitere Ableitungen geläufig sind, z.B. „konfuzianisch“ zu „Konfuzianismus“. Insofern unterscheidet sich dieser Überblick von einem entsprechenden für das Englische, in dem Cannon (1988) auch sämtliche Ableitungen und Komposita verzeichnet, die ein Wort chinesischer Herkunft enthalten, darunter eine ganze Reihe von Wörtern mit „Tee“ als Konstituente. Auch für das Deutsche kann man „Tee“-Komposita wie „Gesundheitstee“, „Magentee“, „Teehändler“, „Teekontor“, „Teeladen“, „Teestube“, „Teezeit“ aufführen, die also ebenfalls eine Konstituente chinesischer Herkunft enthalten. Der Unterschied zwischen Cannons Liste und den deutschen Sinismen ist also deutlich geringer, als er hier zahlenmäßig erscheint.

Nach diesen Vorgaben können folgende Sinismen genannt werden; einige nicht ganz sichere Herkunftsangaben sind durch Fragezeichen gekennzeichnet (s. Tabelle 1):

Tabelle 1
Sinismen im Deutschen

Wort	Jahr-hundert	Bedeutungshinweise	Herkunft
Bonze	16.	buddhist. Priester, Funktionär	frz.–portug.– jap. – chin.
Chan		chines. Bezeichnung für Zen	chin.
Chang		chines. Längenmaß (früher)	chin.
Chanmalerei		Malerei des Chan	chin.
chin-chin		Prost!	engl. - chin.
Chopsuey		Speise	engl. - chin.
Chow-Chow		Hunderasse	engl. - chin.
Chow-Mein		Gericht	chin.
Daimio/ Daimyo		jap. Fürst	chin. – jap.
Dschunke	16.	großes Schiff	engl.-portug.-malai.-chin.

Falun Gong		Schule des Buddhismus	chin.
Fen		kleinste Währungseinheit	chin.
Feng-Shui	20.	Harmon. Lebensgestaltung	chin.
Gehirnwäsche		Massive Beeinflussung des Denkens	engl. "brainstorming": LÜ zu chin. "Hsi-nao"
Ginkgo/ Ginko		Baumart	jap. - chin.?
Ginseng		Pflanze	chin.
Hienfong-Essenz		Hausmittel	chin.
Honanseide		Seidengewebe	chin. (Toponym)
Hong		Gilde; Warenhaus	chin.
Jan-shau-Kultur		jungsteinzeitliche Kultur	chin. (Toponym)
japanisch		von: Japan	chin.
Kalanchoe		Pflanze	frz. - chin.?
Kang		Halsbrett; Schlafbank	chin. (dialektal)
Kaoliang		Hirse	chin.
Kaolin		Ton	frz. - chin. (Toponym)
Ketchup	20.	Soße	engl. – chin.
Kin		Saiteninstrument	chin.
King		Schlaginstrument	chin.
Kombucha		Tee	jap.? - chin.?
Konfuzianismus		Lehre des Konfuzius	chin. (Eigenname)
Kotau	20.	tiefe Verbeugung	engl. - chin.
Kou		Hafen	chin. (Teil v. Toponym)
Kumquat		Orange	engl. - chin. (kanton.)
Kung-Fu		Selbstverteidigung	engl. - chin.
Kuomintang	20.	Partei in Taiwan	chin.
Langschan		Fleischhuhn	chin. (Toponym)
Li		Maßeinheit: Länge, Gewicht	chin.
Limequat		Frucht	chin. „kumquat“
Litchi/ Litschi		Frucht	chin.
Lohan		buddhist. Heiliger	chin. - sanskr.
Loquat		Rosengewächs	chin.
Mah-Jongg		Spiel	engl. – chin.
Maoismus	20.	politische Ideologie	chin. (Eigenname)
Nanking		kräftiges Gewebe	chin. (Toponym)
Orange-Pekoe		Ind. Teesorte	engl. - chin.
Packfong		Legierung	engl. – chin. (kanton.)
Pailou		Ehrentor	chin.
Papiertiger		nur scheinbar starke Person	engl. LÜ chin. "zhilaohu"
Pekinese		Hunderasse	chin. (Toponym)
Pekingente		Gericht; Mastente	chin. (Toponym)
Pekingmensch		Sinanthropus	chin. (Toponym)
Pekingoper		Theaterform	chin. (Toponym)
Pekoe		Teesorte	engl. - chin.
Petong		Kupferlegierung	chin.
Pidginenglisch		Mischsprache	engl. Pidgin - chin. ? entstellt aus "bussiness"
Pinyin	20.	Transskriptionssystem	chin.

Pipa		Chin. Laute	chin.
Pongé		Japanseide	frz. - engl. - chin.?
Qi		Lebensenergie	chin.
Qigong		Heilmethode	chin.
Qigongkugel		therapeuthische Kugel	chin.
Renminbi		Währung	chin.
Samisen		Gitarre	jap. – chin.
Sampan		Boot	chin.
Samschu		Reiswein	chin.
Sanhsien		Laute	chin.
schanghaien		betrunken machen und dann gegen den Willen des Betr. anheuern	engl. (chin. Toponym)
Schantung(seide)		Seidengewebe	chin. (Toponym)
Schen/ Scheng		Mundorgel	chin.
Schogun/ Shogun		Feldherrntitel	jap. - chin.
Sen		jap. Münzeinheit	jap. – chin.
Sen		indones. Münzeinheit	indones. – chin.
Sentoku		Legierung	jap. – chin.
Seppuku		Harakiri	jap. – chin.
Soja	18.	Soße	ndl. - jap. - chin.
Souchong		Teesorte	engl. –chin.
Suanpan		Rechenbrett	chin.
Tai-Chi		Urgrund des Seins; Schattenboxen	chin.
Tai-Chi-Chuan		Schattenboxen	chin.
Tael		alte chines. Münzeinheit	chin.?
Taifun	19.	Wirbelsturm	engl. –chin. (kanton.)
Taikonaut	20.	Weltraumfahrer	chin.
Taipan		Leiter eines ausländischen Unternehmens	chin.
Taiping		Idealzustand	chin.
Tangram		Spiel	engl. - chin.?
Tao		vollkommenes Sein	chin.
Taoismus		Religion	chin.
Tao-Te-King		heilige Schrift des Taoismus	chin.
Tee	17.	Getränk	niederl. - malai. - chin. (Fukien)
Tein, Thein		Inhaltsstoff in Teeblättern	frz. – chin.
Tofu		Nahrungsmittel	jap. - chin.
Trepang		Seegurke (chin. Nahrungsmittel)	engl. – malai. – chin.?
Triade		Geheimgesellschaft	engl. LÜ zu chin. "Gesellschaft der dreifachen Einheit"
Tschan		chin. Buddhismus	chin. – sankr.
Tschekiang		Lammfell	chin. (Toponym)
Tsjao		Münze	chin.
Tungbaum		Lackbaum	chin.

Wok		Kochtopf	chin. (kanton.)
Yamen		Pallast des Siegelbewahrers	chin.
Yang		Männl. Prinzip	chin.
Yangshaokultur		jungsteinzeitl. Kultur	chin. (Toponym)
Yen		Währungseinheit	jap. - chin.
Yin		weibl. Prinzip	chin.
Yinghi		Schattenspiel	chin.
Yuan		Währungseinheit	chin.
Zen		jap. Buddhismus	jap. - chin. - sanskr.

„Bonze“ und „japanisch“ wurden von Cannon (1988: 12, 13) als Wörter mit aus dem Chinesischen stammendem Wortkern übernommen, obwohl in den deutschen Quellen diese Herkunftsangabe fehlt. Mit „Toponym“ wird darauf verwiesen, dass die Bezeichnung auf ein Toponym zurückgeht.

Tabelle 1 listet insgesamt 106 Lexeme auf, die aus dem Chinesischen als Herkunfts- oder Vermittlersprache stammen. Leider können nur 12 davon datiert werden, zu wenig, um auf dieser schwachen Grundlage den Trend des Entlehnungsprozesses zu bestimmen. Darauf wird daher vorerst verzichtet. Es deutet sich aber an, dass die jüngsten Entlehnungen überwiegen. Stattdessen soll ein Blick auf das Englische ein Bild davon vermitteln, wie die Entlehnungen sich in einer europäischen Sprache entwickeln.

Entlehnungen aus dem Chinesischen ins Englische

Die Daten für die Entlehnungen ins Englische lassen sich aus Cannon (1988: 4f.) erstellen. Mit 553 erwiesen sich rund die Hälfte der bekannten Entlehnungen als datierbar. Während die Übernahme von Sinismen erst im 16. Jahrhundert einsetzt, datiert Cannon (1988: 6) das Wort „galingale“ (eine Pflanze, Wurzel) auf das Jahr 1000; es ist hier als Beleg für den Zeitraum bis 1549 eingesetzt. Der chronologische Verlauf stellt sich dann wie folgt dar (s. Tabelle 2):

Tabelle 2
Sinismen im Englischen

Zeitraum	*t*	beobachtet	kumuliert	berechnet
< 1549	1	1	1	6.32
1550-1599	2	6	7	11.50
1600-1649	3	10	17	20.84
1650-1699	4	22	39	37.51
1700-1749	5	42	81	66.73
1750-1799	6	35	116	116.32
1800-1849	7	63	179	196.02
1850-1899	8	148	327	313.64
1900-1949	9	135	462	466.96
1950-1976	9.5	91	553	551.81
a = 328.0965 b = 0.6029 c = 1141.3668 D = 0.9979				

a, *b* und *c* sind die Parameter des Modells; *c* gibt den Zielwert an, auf den nach der Berechnung der Prozess hinausläuft. *D* ist der Determinationskoeffizient, der höchstens den Wert 1 erreichen kann.

Dieser Entlehnungsprozess verläuft also gemäß dem Piotrowski-Gesetz, wie Tabelle und Abbildung 1 belegen. Der Determinationskoeffizient D = 0.9979 signalisiert, dass das Modell hervorragend geeignet ist, um die Entwicklung der Sinismen im Englischen in ihrem Verlauf zu erfassen.

Ein auffallender Unterschied zu vielen anderen Entlehnungsprozessen ist darin zu sehen, dass die Übernahme von Sinismen sich offenbar keineswegs ihrem Ende nähert. Ähnlich wie die Anglizismen im Deutschen nehmen die Entlehnungen aus dem Chinesischen ins Englische in hohem Maße zu (Best 2006: 112f.; Körner 2004: 36). Ein Ende des Prozesses deutet sich nicht an.

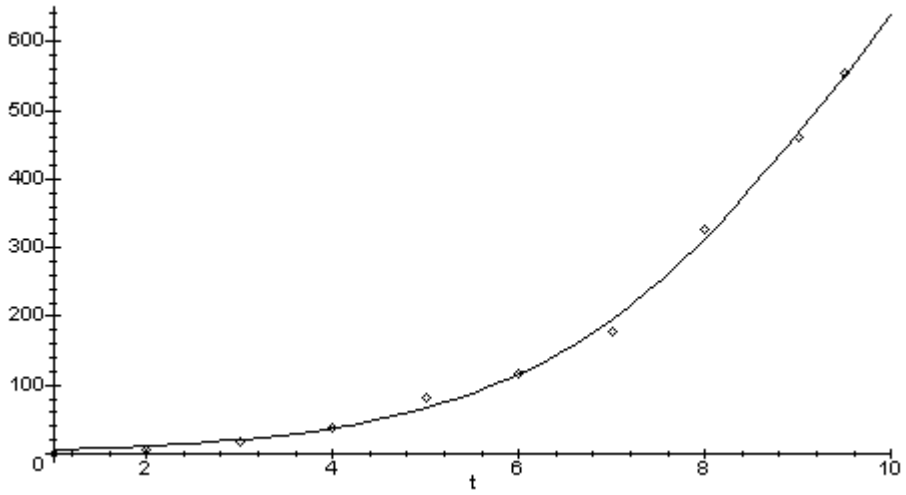


Abb.1 Die Entwicklung der Sinismen im Englischen (In dieser Graphik steht t = 1 für die Entlehnungen bis 1549, t = 2 für den Zeitraum 1550-1599; etc.)

Zusammenfassung und Perspektive

Die Beobachtungen zu den Sinismen im Deutschen und Englischen zeigen, dass der Einfluss des Chinesischen derzeit offenbar deutlich zunimmt. Auch wenn die gesicherten Daten für das Deutsche noch zu gering sind, kann vermutet werden, dass die Entlehnungen in beiden Sprachen einen ähnlichen Verlauf nehmen, wenn auch nicht unbedingt auf dem gleichen zahlenmäßigen Niveau. Auch im Deutschen überwiegen jedoch die Übernahmen im 20. Jahrhundert. Manche Aspekte chinesischer Kultur und Wissenschaft üben in Europa eine beachtliche Faszination aus; mit der zunehmenden Bedeutung Chinas in der Welt wird dieser Einfluss wohl auch nicht abnehmen, was sich vermutlich auch in Zukunft in unseren Sprachen bemerkbar machen wird. Der erkennbare Trend im Englischen ist ein überzeugender Hinweis darauf.

Literatur

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

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

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

Cannon, Garland (1988). Chinese borrowings in English. *American Speech 63, 3-33.*

Duden. *Das große Fremdwörterbuch. Herkunft und Bedeutung der Fremdwörter* ([4]2007). 4., aktualisierte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

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

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

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

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

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

Strauss, Udo, Fan, Fengxiang, & Altmann, Gabriel (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.

Verwendete Software

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

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

Glottometrics 17, 2008, 94-108

On the regularity of diversification in language

Ioan-Iovitz Popescu, Bucharest[1]
Gabriel Altmann, Lüdenscheid

Abstract. Diversification seems to abide by a strong regularity which does not depend on language but rather on the language level (category). The regularity is evidenced using a relationship between some indicators of the rank-frequency distribution.

Keywords: *Diversification, language levels, h-point, arc length*

1. Introduction

Diversification is a process existing in all language phenomena. It gives rise to variants, dialectal forms, new languages, style, polysemy, polymorphy, synonymy, etc. An overview of possible diversifications can be found in Rothe (1991b). Certain types of diversification result in very regular frequency distributions, and some of the diversifications have already been scrutinized (cf. Rothe 1991; Altmann 2005; Best 2008). Diversification has phylogenetic, historical, synchronic, psychological, somatic, social and other causes. Some boundary conditions – whose identity is not yet known – force the data to abide by different distributions either in different phenomena in the same language or in the same phenomenon in different languages. The background is not yet known but the phenomenon itself is real. A word can have different meanings, an object can have different names, a speech sound produced by the same person can have different pronunciations, etc. Thus it is quite natural to search for an entity which is more constant and characteristic for the given phenomenon than the varying distributions.

In previous articles (cf. Popescu, Mačutek, Altmann 2008; Fan, Popescu, Altmann 2008) it has been stated that in rank frequency distributions of whatever kind there is a regularity leading to a very stable relationship between inventory (= highest rank, vocabulary, R), the frequency of the most frequent entity (f_1), arc length (L) and the h-point. It has been shown that

$$L = R + f_1 - ch \tag{1}$$

where all variables can be computed from the data. The variables L, R, f_1 and h lie in an interval whose right boundary is infinity, however, data of the same type tend to a mean value of c, a kind of constant whose investigation may be of interest for quantitative linguistics.[2]

In word frequency analysis using 100 texts in 20 languages it has been observed that the average c regardless of language or types of text varies in a small interval around $c = 1.30$. For the diversification of meanings of 165 randomly sampled English words it was found $c = 1.47$. Using the same method and the data of Meuser, Schütte & Stremme (2008) on the rank-frequency distributions of plural allomorphs of nouns in German one can obtain $c = 1.36$. This

[1] Address correspondence to: iovitzu@gmail.com.

[2] Further developments and verification of the relationship (1) will be presented in the coming book entitled *New Aspects of Word Frequencies*, Chapter 5 Arc length (Popescu, Mačutek, Altmann 2008).

automatically evokes the question concerning the background of the quantity *c*. The coincidence testifies to the fact that (1) rank-frequency distributions in language are a law-like phenomenon developing very regularly as a result of a self-organizing process, and (2) the parameter *c* either varies randomly or expresses a specific character of the given phenomenon. In order to make a further step in this direction, we analyzed all available data on diversification.

If the sample is too large but the inventory is small so that the greatest rank is still smaller than the smallest frequency, i.e. $R = r_{max} < f_R$, the parameter *h* cannot be computed as usual. Hence we modify the computation of *h* subtracting $(f_R - 1)$ from all frequencies. For example a sequence of frequencies

r	1	2	3	4	5
f_r	100	90	80	70	60

will be changed to

1	2	3	4	5
(100-59)	(90-59)	(80-59)	(70-59)	(60-59)
41	31	21	11	1

in which *h* can be computed by interpolation between ranks 4 and 5. Then, in general we compute *h* as follows:

$$(2) \qquad h = \begin{cases} r & \textit{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \textit{if there is no } r = f_r \end{cases}$$

For computing (2) we take such values that $f_1 > r$ and $f_2 < r + 1$. . If in (2) $r_2 = r_1 + 1$, the formula can be slightly simplified. In our example $r_1 = 4$, $f_1 = 11$, $r_2 = 5$, $f_2 = 1$. The systematic shifting of frequencies does not change the arc length. For computing *c* we will generalize (1) in the form $L = R + f_{max} - f_{min} + 1 - ch$, hence

$$(3) \qquad c = \frac{R + f_{\max} - f_{\min} + 1 - L}{h}$$

which holds even in case that the smallest frequency f_{min} was 1.

The arc length *L* is computed empirically as the sum of Euclidean distances between neighbouring frequencies, that is

$$(4) \qquad L = \sum_{r=1}^{R-1} [(f_r - f_{r+1})^2 + 1]^{1/2}$$

2. Analysis of language levels

The data at our disposal concern very different phenomena in a number of languages, so that one can get only a first sight of the phenomenon. Hence, all results must be taken cum grano salis and further data must be collected in order to corroborate whatever hypothesis thereon. We prefer to speak about language phenomena rather than about language levels because the present investigation only partially corresponds with our rather intuitive conception of language level.

2.1. Let us begin with s**ounds, phonemes** and **letters** which do not have their own meaning. In order to avoid the well known problems associated with the identification of these units, we took published data from different languages. All Russian data were taken from Grzybek, Kelih (2003). The English data were obtained from texts available on the Internet (http://www.gutenberg.org/browse/scores/top). Fry (1947) was considered on historical grounds.

The data and the results of computation are presented in Table 1. As can be seen, the value of *c* is stable both historically – as shown in Russian data from different years – and cross-linguistically, hence, it is no typological feature but, perhaps, characterizes the rank-frequency distribution on a certain level.

Table 1
Computing *c* for sounds, phonemes and letters in different languages

Source	**Data**	***R***	***h***	***L***	***c***
Ch. Dickens, David Copperfield, letters	181444, 132988, 122911, 116398, 108978, 103014, 91010, 90908, 86432, 70508, 56873, 47414, 42518, 39219, 33828, 33787, 32638, 31522, 25472, 23018, 13899, 13502, 2118, 1442, 1420, 267;	26	25.98	181177	1.04
Ch. Dickens, Great Expectations, letters	91775, 68686, 62912, 59877, 54472, 52710, 48167, 45097, 40849, 36636, 28012, 22549, 21733, 20231, 16938, 16443, 16067, 15743, 12912, 12167, 7535, 6718, 1669, 942, 702,209;	26	25.95	91566	1.04
Ch. Dickens, A Christmas Carol, letters	14914, 10914, 9720, 9336, 8396, 8328, 7973, 7939, 7058, 5694, 4571, 3345, 3104, 3052, 2986, 2849, 2445, 2308, 2126, 1949, 1032, 1031, 133, 112, 97, 86;	26	24.19	14829	1.07
J. Joyce, Ulysses, letters	141465, 100184, 93129, 91404, 81407, 80139, 76915, 72550, 69853, 55053, 49093, 33272, 31534, 29894, 27791, 26638, 26164, 24251, 22441, 21150, 12067, 9755, 2317, 1436, 1332, 1077;	26	25.9	140388	1.04
C. Doyle, Sherlock Holmes, letters	53034, 38947, 35112, 33454, 30112, 29020, 28625, 27146, 24487, 18510, 17131, 13067, 11781, 11264, 10487, 9424, 8966, 7884, 6792, 6350, 4438, 3541, 545, 452, 426, 148,	26	25.91	52886	1.04
M. Twain, Huckleberry Finn, letters	47144, 40770, 35504, 35298, 31709, 27024, 25774, 24378, 23176, 19016, 16926, 13403, 13057, 10309, 9964, 9950, 7687, 7497, 7141, 5522, 5497, 2818, 1122, 422, 182, 178;	26	24.92	46966	1.08
J. Milton, Paradise Lost, letters	42728, 32212, 27245, 26224, 24881, 24850, 24139, 23998, 23094, 16754, 15249, 10900, 8871, 8710, 8650, 7670, 7623, 6133, 5352, 4415, 3886, 1961, 459, 456, 252, 176;	26	25.67	42552	1.05
H.G. Wells, The War of the Worlds, letters	33398, 25669, 22438, 19236, 18626, 18015, 16431, 16165, 15687, 12846, 10198, 7029, 6851, 6393, 6181, 6166, 6048, 4752, 4675, 4010, 2358, 2020, 362, 185, 179, 105;	26	25.67	33293	1.05
J. Swift, Gulliver's Travels, letters	58078, 41943, 35407, 35270, 31158, 30812, 28218, 27759, 27511, 19671, 16527, 13270, 13164, 12106, 11044, 10255, 9079, 8408, 8241, 7150, 4984, 2569, 821, 593, 507, 145;	26	25.93	57933	1.04
E. Bronte, Wuthering Heights, letters	63773, 42950, 38734, 36689, 35736, 35653, 32685, 30452, 29229, 24017, 20661, 14782, 13199, 11788, 10776, 10673, 10543, 10463, 7759, 6864, 4568, 3899, 820, 550, 472, 198;	26	25.91	63575	1.04
Ch. Bronte, Jane Eyre, letters	100613, 67050, 62795, 60534, 56018, 54233, 50085, 47366, 45855, 37427, 32463, 23473, 22261, 18762, 18688, 17288, 16875, 15077, 12128, 11147, 7619, 6028, 1276, 1222, 946, 328;	26	25.96	100285	1.04
B. Stoker, Dracula, letters	79310, 58139, 52342, 50340, 43601, 43213, 42610, 39485, 34957, 28536, 26116, 18061, 17991, 17758, 13993, 13517, 12675, 12672, 9159, 8986, 6201, 5872, 815, 781, 625, 351;	26	25.91	78959	1.04

English sounds (Fry 1947)	51830, 34260, 33922, 31934, 31009, 27373, 23069, 22453, 21275, 19972, 15553, 14823, 11603,11312, 10907, 10234, 9181, 8839, 7960, 7441, 7124, 6721, 6685, 6239, 6117, 6007, 4794, 4730,4600, 4215, 4174, 3869, 3699, 3560, 3095, 2672, 2179, 1977, 1602, 1053, 965, 788, 596, 334;	44	43.84	51496	1.03
Finnish letters (Pääkkönen 1994)	296538, 265007, 243269, 215911, 204445, 195675, 141553, 132990, 130545, 126164, 114254, 82272, 62780, 57822, 47591, 45503, 44668, 43282, 21070, 12188, 3146, 1593, 1539,1041, 307, 30, 25;	27	25.93	296513	1.08
Georgian phonemes (Job 1974)	2064,1446,1072,821,774,744,665,595,578,548,490,428,364 ,349,263,218,213,211,168,124,121,114,112,103,88,79,63,5 7,48,28,26,17,9;	33	29.52	2057	1.08
German sounds (Meier 1964; Best 2004/2005)	10275,8666,8623,7284,4476,4472,3962,3894,2864,2856,24 64,2458,2449,2298,2291,2207,2164,2145,2114,2109,1751, 1724,1568,1534,1288,1126,980,971,852,765,721,663, 441, 325, 297,250,242,226,197,102,92,68,66,38,6,1;	46	43.79	10275	1.05
Hawaiian letters (Schulze 1974)	5305, 2395, 2233, 1752, 1400, 1274, 1272, 1164, 1074, 1005, 394, 110, 80;	13	12.61	5225	1.11
Sea Dayak letters (Rademacher 1974)	4428, 1914, 1847, 1254, 1156, 1036, 976, 958, 871, 752, 727, 670, 654, 625, 576, 575, 417, 325, 200, 23, 15;	21	19.94	4414	1.05
Slovenian letters (Grzybek, Kelih 2006)	32036, 31891, 31122, 27150, 22905, 16088, 16084, 15221, 14668, 14043, 13034, 10517, 10514, 10216, 9568, 7446, 6413, 5361, 5055, 4608, 2606, 2554, 2463, 1675, 497;	25	24.98	31539	1.04
Slovak letters (Grzybek, Kelih 2006)	14193, 13772, 12700, 9285, 8323, 7099, 6562, 6534, 6163, 6091,5731,5659,5229,4121,3845,3376,3135,3014,2784, 2676,2408,2262,1825, 1685, 1613, 1465, 1395, 1294, 1072, 947, 719, 402, 346, 297, 270, 253, 172, 131, 47, 27, 10, 3;	42	39.29	14190	1.09
Serbian letters (Grzybek, Kelih 2006)	885, 845, 845, 811, 548, 512, 451, 445, 438, 370, 339, 318, 233, 225, 206, 206, 183, 170, 112, 81, 67, 66, 56, 56, 48, 42, 35, 17, 14;	29	26.38	875	0.99
Russian letters (Ol´chin 1907)	2460, 2375, 1670, 1660, 1656, 1502, 1231, 1081, 948, 870, 851, 817, 738, 646, 516, 419, 416, 357, 337, 317, 229, 219, 201, 198, 178, 175, 100, 87, 50;	29	28.26	2411	1.03
Russian letters (Proskurin 1933)	110020, 87161, 75060, 74400, 64932, 64510, 54942, 47658, 45276, 42022, 33662, 31192, 30197, 28019, 24808, 21874, 19702, 17546, 17281, 17029, 15908, 14926, 11831, 10715,9737, 7271, 6831, 4481, 4449, 3165, 1884, 382, 331;	33	32.38	109689	1.05
Russian letters (Kalinina 1968)	11376, 8907, 7852, 7338, 7020, 6889, 5498, 5116, 4227, 4104, 3358, 3072, 3047, 2641, 2302, 1919, 1915, 1752, 1563, 1364, 1256, 1210, 1200, 1032, 789, 753, 692, 477, 460, 449, 422;	31	29.83	10954	1.07
Russian letters Grigor´ev 1980a)	5678, 4206, 3979, 3349, 3112, 2983, 2511, 2334, 2174, 2091, 1981, 1555, 1527, 1493, 1294, 1068, 1052, 990, 968, 923, 798, 780, 706, 557, 512, 480, 341, 170, 168, 162, 42, 16;	32	30.97	5663	1.03
Russian letters (Grigor´ev 1980b)	11410, 8610, 8002, 6536, 6097, 5926, 5072, 4674, 4492, 4157, 4140, 3098, 3095, 2977, 2488, 2092, 2090, 1981, 1939, 1912, 1611, 1490, 1373, 1130, 1012, 857, 685, 323, 310, 304, 81, 22;	32	31.48	11389	1.02
Russian letters (Dietze 1982)	44172, 42024, 35662, 33967, 29877, 27447, 26034, 22279, 17586, 14613, 14189, 13890, 12736, 11079, 9893, 8632, 8413, 7000, 6464, 6005, 5390, 4852, 4716, 4491, 4389, 3912, 2904, 2537, 1670, 1224, 1054, 156;	32	31.97	44016	1.03

The mean $\overline{c}$ = 1.0489, the standard deviation of the individual values is s_c = 0.0245 and the standard deviation of the mean is $s_{\overline{c}}$ = 0.0047. For the sake of simplicity we assume normality and set up 95% confidence intervals for individual values as $\overline{c} \pm 1.96 s_c$ and for the mean

$\overline{c} \pm 1.96\, s_{\overline{c}}$. Hence for individual values we obtain <1.0009, 1.0969>, that for the mean <1.0397, 1.0581>.

2.2. Rhythmic units are based mostly on suprasegmental features and have nothing common either with grammar or with semantics. We use here line patterns in Latin, Greek and German hexameter and distych. The data were taken from Drobisch (1866, 1872, 1875, 1868a,b) and presented for other purposes in Best (2008). Since the last two feet in the verse are identical, it is sufficient to consider the combinations of dactyls (D) and spondees (S) in the first four positions. One obtains 16 patterns like SSSS, SSSD, SSDS, SDSS,… The authors may differ in the use of individual patterns whose identity is here irrelevant, we consider only the rank order of patterns. All data are samples of uninterrupted sequences from the works of the given authors. We obtained from Drobisch the data whose evaluation is presented in Table 2.

Table 2
Hexameter patterns in Latin, Greek and German
(Drobisch 1866, 1868a,b, 1872, 1985)

Text	**Pattern sequences**	***R***	***h***	***L***	***c***
Goethe, “Reinecke Fuchs”	204,181,96,93,86,71,65,46,43,42,38,22,14,7,6,5	16	12.67	201	1.18
Goethe, “Hermann und Dorothea”	200,149,142,104,98,73,63,41,40,35,34,31,25,25,15,11	16	14.09	192	0.99
Goethe, “Elegien”	96,92,72,51,45,37,32,28,19,11,8,6,6,6,5,4	16	9.78	96	1.33
Leibniz, “Epicedium”	59,57,33,31,30,24,20,16,13,12,10,10,10,6,5,2	16	10.33	62	1.16
Klopstock,: “Messias”	129,128,125,102,76,66,65,60,60,52,46,45,26,13,6,3	16	13.78	129	1.02
Voss, “Luise”	188,173,164,161,154,135,131,95,70,51,44,35, 32,9	14	13.46	180	1.04
Voss, “Odyssey”	125,123,114,98,96,91,86,67,65,59,53,39,31,16,4,1	16	14.15	126	1.06
Vergil, “Georgica”	84,62,55,49,39,31,31,29,29,19,19,18,16,12,11,9	16	11.00	80	1.09
Vergil, other sample	78,75,57,52,44,38,37,33,30,27,22,20,18,13,12,4	16	13.33	76	1.13
Vergil, “Aeneis“	423,338,325,297,229,191,170,167,167,136,117,106, 102,66,60,58	16	13.87	367	1.08
Vergil, “Bucolica”	107,90,79,64,63,59,57,43,43,40,38,29,27,26,23,21	16	11.70	89	1.20
Horace, another sample	62,53,49,48,44,38,38,36,35,32,30,22,22,20,20,11	16	12.00	56	1.00
Horace, “Satires”	285,228,205,203,174,137,134,115,108,102,93,92,79, 63,48,46	16	14.25	240	1.12
Horace, “Epistulae”	237,198,189,168,147,132,125,124,109,109,97,94,89, 59,54,36	16	15.21	203	1.05
Lucrece, “De rerum natura”	88,72,63,56,51,39,37,36,26,21,17,17,12,10,8,7	16	11.00	84	1.27
Manilius, “Astronomica”	93,67,60,57,48,34,33,30,28,22,22,19,16,11,11,9	16	11.75	88	1.11
Persius, “Satires”	118,96,68,62,48,39,35,35,32,30,27,19,14,12,9,5	16	12.50	116	1.12
Juvenal, “Satires”	85,67,64,51,40,38,35,29,26,25,22,20,19,14,13,12	16	11.00	76	1.27
Lucanus, “Pharsalia”	98,83,59,58,39,33,32,29,28,23,19,15,13,12,11,8	16	11.20	93	1.25
Quintus Ennius, “Fragments”	64,39,39,35,25,24,24,24,23,21,20,20,19,15,12,10	16	11.00	61	0.91
Catull, 2 poems	124,65,55,51,43,25,15,15,8,7,6,5,4,3,3,1	16	8.88	128	1.35
Ovid, “Metamorphoses”	78,76,63,60,57,54,45,33,26,21,13,11,6,6,6,5	16	10.78	77	1.21
Silius Italicus, “Punica”	75,63,54,53,47,47,34,32,28,26,25,24,18,16,11,7	16	12.86	72	1.01
Valerius Flaccus, “Argonautica”	131,75,64,63,54,52,49,30,24,24,22,21,12,11,5,3	16	12.70	131	1.10

Statius, "Thebais"	83,76,57,53,43,40,39,34,32,24,17,16,14,14,10,8;	16	10.88	78	1.29
Claudian,: "Raptus Proserpinae"	102,83,75,67,51,38,34,30,24,21,14,8,5,3,3,2;	16	11.29	103	1.24
Homer, "Iliad"	350,320,296,196,155,149,145,78,76,73,56,28, 22,21,19,8;	16	14.00	343	1.14
Homer, "Odyssey"	410,323,277,185,176,161,149,92,82,71,65,35,34, 20, 18, 5;	16	14.67	406	1.09
Theokrit, "1st Idyll"	31,21,13,11,10,9,7,5,5,5,4,2,1;	13	7.00	35	1.29
Theognis, "Elegic poems"	117,99,78,66,38,36,34,28,26,25,23,21,7,4,3,2;	16	12.53	118	1.12

As can be seen, $\overline{c}$ = 1.1407, s_c = 0.1106 and $s_{\overline{c}}$ = 0.0202. The 95% confidence interval for individual values is <0.923989, 1.3575> and for the mean <1.1011, 1.1803>.

2.3. Word classes defined in classical Latin manner have been published by different authors. The classification of words in classes can be performed in different ways resulting in dozens of classes, but it is not our aim to compare them. We simply study the forming of the rank-frequency distribution in a possible classification. The data and the results are presented in Table 3. The data were taken from the following sources: Latin, German, Chinese (Schweers, Zhu 1991), Polish (Sambor 1989), German (Best 1994), Portuguese, Brazilian Portuguese (Ziegler 2001).

Table 3
Computation of *c* for word classes

Language	**Data**	***R***	***h***	***L***	***c***
Latin	347, 173, 142, 98, 93, 59, 40, 39, 9;	9	8.74	339	1.08
German	192, 161, 153, 112, 111, 104, 97, 70;	8	7.75	123	1.07
Chinese	247, 228, 140, 133, 107, 81, 55, 27;	8	7.95	220	1.10
Polish	144188, 79995, 71988, 56812, 33605, 31833, 21428, 18757, 8076, 650;	10	8.73	143538	1.26
German (Best)	2032, 1939, 1532, 1338, 1179, 974, 914, 761;	10	10.00	1271	1.10
Portuguese	2586, 1607, 949, 819, 776, 680, 478, 440, 352;	9	8.91	2234	1.12
Brazilian Portuguese	2930, 2265, 1743, 1708, 1602, 1040, 936, 394;	9	8.00	2536	1.25

This way of determining word classes, though not the only possible one, yields a very compact result. In Portuguese numerals were considered a separate word class. In German (Best 1994) the unique interjection has been omitted. As can be seen, the text size (*N*) does not play any role. There is no difference between languages, hence the quantity *c* cannot be used for typological purposes. It is a matter of the given phenomenon. For word classes we obtain mean $\overline{c}$ = 1.1385, the standard deviation s_c = 0.0798, thus the values lie in the 95% interval <0.9821, 1.2949> and the mean in <1.0794, 1.1803>

2.4. In two cases the **allomorphs of the German plural** have been studied. Meuser, Schütte and Stremme (2008) analyzed 21 short stories by Wolfdietrich Schnurre, and Brüers and Heeren (2004) the individual letters of Heinrich von Kleist. The ranks differ in all texts, sometimes not all allomorphs are used in a unique text, but the distributions display the same tendencies. The data and the results are presented in Table 4.

Table 4
Computing *c* for the allomorphs of the German plural
(Texts I: Meuser, Schütte, Stremme (2008); Texts II: Brüers, Heeren (2004))

Texts I	**Data**	***R***	***h***	***L***	***c***
1	9,9,7,6,5,5,3,1	8	5.00	11.54	1.09
2	28,16,14,13,8,3	6	5.17	25.89	1.18
3	17,7,7,6,4,3,2,2	8	4.66	18.53	1.17
4	20,14,4,4,3,2,1	7	4.00	21.37	1.41
5	11,8,7,6,5,2,1	7	5.00	11.98	1.20
6	20,7,5,3,2,2,2,2	8	3.33	21.92	1.53
7	21,18,10,8,5,3,2,2	8	4.75	21.27	1.42
8	59,33,20,18,18,12,10,7	8	6.00	53.77	1.20
9	12,9,8,8,4,3,3,1	8	4.80	14.35	1.18
10	51,25,14,13,5,3,2,1	8	5.00	51.61	1.48
11	20,13,10,9,8,3,1,1	8	5.50	21.40	1.20
12	20,7,7,6,5,4,4	7	3.50	19.28	1.35
13	18,9,9,7,6,5,1	7	5.50	19.24	1.05
14	125,80,63,43,42,13,8,8,4	9	6.67	122.72	1.24
15	10,10,4,3,2,2,2,1,1	9	3.50	14.33	1.33
16	39,27,20,14,12,11,4,1,1	9	6.62	40.08	1.20
17	18,13,11,4,2,1	6	4.00	18.06	1.48
18	18,12,8,8,6,4	6	4.33	15.68	1.23
19	33,18,15,13,8,7,7,3	8	5.50	32.07	1.26
20	63,19,18,17,14,9,8,2,2	9	7.00	63.60	1.06
21	11,11,3,3,3,1	6	3.00	13.30	1.23
Texts II	**Data**	***R***	***h***	***L***	***c***
1	17,7,6,5,3,1	6	4.33	17.35	1.30
2	8,6,4,3,2,1,1	7	3.50	9.71	1.51
3	10,3,2,1,1,1,1	7	2.50	12.90	1.64
4	6,3,1	3	2.33	5.40	1.54
5	12,12,4,4,2,1	6	4.00	13.71	1.07
6	6,3,2,1	4	2.50	5.99	1.6
7	5,4,2,2	4	2.33	4.65	1.44
8	9,3,2,1	4	2.50	8.91	1.64
9	11,7,5,4,4,2	6	3.50	11.01	1.43
10	6,4,4,2,1	5	3.33	6.89	1.23
11	5,2,2,2,2,2	6	1.75	7.16	1.62
12	16,10,4,3,2,2,1	7	3.50	17.41	1.6
13	8,6,3,2,1	5	3.00	8.23	1.59
14	4,4,3,2,1	5	3.00	5.24	1.25
15	7,5,2,1	4	2.75	6.81	1.52
16	6,4,2,2,2	5	2.33	6.47	1.51
17	7,6,4,3	4	2.67	5.06	1.48
18	18,17,10,5,5,4,4,2	8	4.00	19.23	1.44
19	3,3,3,3,1	5	3.00	5.23	0.92
20	7,2,1,1,1	5	2.00	8.51	1.74
21	23,10,8,8,6,4	6	4.33	20.75	1.21

The mean $\overline{c}$ in the first case is 1.2616, jn the second 1.4435, jointly $\overline{c}$ = 1.3525. The relatively long range of resulting values indicates that here perhaps, the contents of the given texts play a certain role. The standard deviation is s_c = 0.1966, hence the 95% confidence intervals are: c ∈ <0.9672, 1.73785> and $\overline{c}$ ∈ <1.2930, 1.4120>.

2.5. Prepositions, postpositions and conjunctions are auxiliaries underlying strong diversification because they occur frequently and in many contexts. Here specimens from 5 languages are presented in Table 5.

Table 5
Rank-frequency distributions of auxiliaries

Auxiliary	**Data**	***R***	***h***	***L***	***c***
Japanese: postposition *ni* (Roos 1991)	40,32,31,27,14,11,6,6,4,4,2,1;	12	6.83	42.79	1.35
German: particle/preposition *von* (Best 1991)	54,38,21,21,19,16,15,13,12,11,11,10, 9,9,8,8,7,7,6,6,5,5,5,5,4,4,4,3,3,3,3, 3,3,3,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1, 1,1,1;	53	11.00	93.08	1.27
German: preposition *auf* (Th.Mann) Fuchs (1991)	24,12,12,6,6,6,4,3,3,2,2,2,2,2,2,2, 2,2,1,1,1,1,1,1,1,1,1;	27	6.00	44.60	1.07
German: preposition *auf* (C. Wolf) Fuchs (1991)	312,152,123,41,38,34,33,30,30,26,25, 24,23,20,12,11,10,10,10,9,9,8,7,7,6,6, 6,6,5,5,5,5,4,3,3,3,2,2,2,2,2,2,1,1,1,1, 1,1,1,1,1,1,1,1,1;	54	14.67	347.03	1.29
English: preposition *in* Hennern (1991)	51,49,14,14,14,10,8,7,6,5,5,5,5,4,4, 4,4,4,4,4,3,2,2,1,1,1,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1;	43	7.50	84.51	1.27
Polish: preposition *w* Hammerl,Sambor(1991)	199,100,55,40,21,21,20,15,13,8,7,1;	12	9.67	200.42	1.09
Russian: conjunction *no* (Kuße 1991)	19,16,13,10,9,8,8,7,6,5,4,3,3,1,1,1;	16	7.50	25.62	1.25
Russian: conjunction *a* Kuße (1991).	22,11,8,6,6,5,4,3,3,2,2,2,2,1,1,1,1,1;	18	5.50	32.51	1.36

The mean $\overline{c}$ = 1.2432, the standard deviation is s_c = 0.1083, and the 95% confidence intervals are: c ∈ <1.0309, 1.4555> and $\overline{c}$ ∈ <1.1682, 1.3182>

2.6. Semantic diversification of prefixes and suffixes. Unfortunately, we have only data from four languages at our disposal, namely German: Rothe (1989), Altmann, Best, Kind (1987); Middle High German: Kaliuščenko (1988); Slovak: Nemcová (1991, 2007) and Hungarian: Beöthy, Altmann (1984a,b). The results are presented in Table 6.

Table 6
The quantity *c* with prefixes and suffixes

Affix	**Data**	***R***	***h***	***L***	***c***
Hungarian *föl-*	11,7,7,6,5,4,3,3,3,3;	10	4.00	13.77	1.31
Hunarian *el-*	83,9,3,2,2,1,1,1,1;	9	3.00	86.92	1.69

Hungarian *be-*	20,11,10,7,5,3,3,3,2,1,1,1,1;	13	5.00	25.83	1.43
Hungarian *ki-*	12,10,8,8,5,4,3,3,2,2,2;	11	4.75	15.87	1.29
Hungarian *meg-*	107,8,8,5,5,4,2,1,1;	9	5.00	110.23	1.15
German *ab-*	16,7,3,3,2,1,1,1,1,1,1;	11	3.00	22.01	1.66
German *aus-*	24,10,10,9,5,4,4,3,3,2,2,1;	12	5.00	29.23	1.35
German *be-*	86,13,10,8,6,3,2,2,2,1,1,1,1,1,1,1;	16	5.25	94.63	1.40
German *ein-*	18,4,4,3,2,2,2,1,1;	9	3.50	22.28	1.35
German *ent-*	71,12,3,3,3,1;	6	3.00	72.30	1.57
German *ver-*	42,40,32,17,9,7,4,4,4,3,2,1,1;	13	6.25	46.03	1.43
Middle High German *be-*	36,4,3,3,2,2,2,2,2,1,1,1,1;	13	3.00	44.26	1.58
Middle High German *ent-*	46,5,3,2,1;	5	3.00	46.08	1.64
Middle High German *ver-*	14,10,4,3,3,2,2,2,2,1,1,1,1,1,1;	15	3.50	23.45	1.59
Slovak *do-*	22.20,6,3;	4	3.25	19.43	1.41
Slovak *na-*	30,23,17,16,10,7;	6	4.86	23.81	1.27
Slovak *o-*	17,15,15,11,4;	5	4.50	13.88	1.14
Slovak *ob-*	19,8,4;	3	2.60	15.17	1.47
Slovak *od-*	25,8,5,4,4,4;	6	2.75	23.61	1.60
Slovak *po-*	59,54,29,24,18;	5	4.43	41.30	1.29
Slovak *pre-*	33,16,13,7;	4	3.57	26.27	1.32
Slovak *pri-*	26,21,7;	3	2.87	19.13	1.35
Slovak *roz-*	26,26,25,22;	4	3.25	5.58	1.05
Slovak *s-/z-*	71,38,18,15,7,5;	6	4.78	66.50	1.36
Slovak *u-*	56,39,33,2;	4	3.91	54.13	1.25
Slovak *vy-*	61,47,33,32,23;	5	4.60	38.54	1.19
Slovak *za-*	77,35,21,8;	4	3.79	69.09	1.30
German *–os/ös*	59,31,20,6,5,4,2,2,2;	9	4.50	60.16	1.52
German *–al/-ell*	93,68,56,46,32,23,20,19,11,7,7,6,6,4,4,1;	16	9.90	96.78	1.23

The preliminary $\overline{c}$ = 1.3861, s_c = 0.1674, so that c ∈ <1.0580,1.7142> and $\overline{c}$ ∈ < 1.3252, 1.4470>.

2.7. Grammatical categories. There is only one case of diversification known, namely Rothe (1991a) in which the functions and the meanings of the German genitive are analyzed. The data are given in Table 7.

Table 7
Diversification of the German genitive

Category	**Data**	***R***	***h***	***L***	***c***
German genitive	47,31,28,27,24,23,21,21,13,9,9,8,8,7,6,6,6,6,5,4, 4,4,3,3,3,3,3,3,3,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1;	55	9.80	88.92	1.33

2.8. In addition to the semantic diversification of word **meanings** in English (Fan, Popescu, Altmann 2008) there is a well scrutinized case of French *et* in *Le petit prince* by A. Saint-Exupéry by U. Rothe (1986), who found 70 senses and the distribution is given in Table 8

Table 8
Diversification of French *et* (Rothe 1986)

	Data	***R***	***h***	***L***	***c***
French *et* (Rothe 1986)	17,17,13,11,9,6,5,5,5,5,4,4,4,4,4,4,3,3,3,3,3,3,2, 2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1;	70	6	78.83	1.36

Adding this result to the 165 cases in English the mean $\overline{c}$ does not change, it is 1.47.

2.9. J. Sambor (1989) studied the distribution of **inflection classes of nouns and verbs** in the Polish frequency dictionary. She took two aspects into account: the number of nouns and verbs belonging to a given class and the frequency of the entire class in the dictionary. In this way she obtained four diversification cases:

Number of nouns belonging to a special inflection class
Number of verbs belonging to a special inflection class
Frequencies of individual noun classes
Frequencies of individual verb classes.

The results are given in Table 9

Table 9
Diversification of Polish inflection classes (Sambor 1989)

	Data	***R***	***h***	***L***	***c***
Nouns, number	2624,2557,1473,1407,1196,1011,729, 569,403,357,273,159,149,56,54,40,32, 31,27,21,20,14,9,8,8,5,4,1;	28	20.50	2627	1.23
Verbs, number	1728,1115,970,456,331,218,203,200,112, 108,103,78,78,63,62,61,56,54,16,9,8,8,6, 5,2,1,1;	27	18.92	1733	1.14
Nouns, frequency	34288,20034,17370,14755,11677,8326,7626, 6475,6121,6021,2553,2121,1263,1081,683, 341,245,175,171,107,84,73,39,38,25, 7,6,4;	28	24.79	34285	1.11
Verbs, frequency	19292,13865,10413,4804,3935,3224,3120, 2485,1859,1184,1107,1084,1011,886,633, 541,529,450,317,251,244,132,57,15,6,1,1;	27	23.79	19292	1.12

Taken all together, we obtain the mean mean $\overline{c}$ = 1.1515, s_c = 0.0530, and $c \in$ <1.0476, 1.2554>, $\overline{c} \in$ <1.0996,1.2034>

2.10. Colour names and their frequencies in several languages were studied by A. Pawlowski (1999). Colour names are a closed and relatively small semantic class of adjectives. It has been observed that the rank-frequencies follow a usual law. Pawlowski took into account only colours, hence the data are not complete. Some colours were not present in the frequency dictionary and their number was given as 0. Since this is not the usual way of counting frequencies, the results must be considered cum grano salis. Nevertheless, one can at least have a tentative look at the data. The computation of *c* is presented in Table 10.

Table 10
Colour names in 10 languages (Pawlowski 1999)

Language	**Data**	***R***	***h***	***L***	***c***
Czech	604,519,416,206,205,158,146,127,96,16,4,3;	12	10.31	601.98	1.17
English	365,203,197,176,143, 116, 55, 48, 23, 13, 12, 7;	12	9.73	358.81	1.25
French (Juilland)	136, 113, 74, 58, 35, 20, 17, 7, 0, 0, 0, 0;	12	8.00	139.40	1.20
French (Engwall)	298, 278, 170, 134, 101, 98, 77, 61, 12, 9, 0, 0;	12	10.00	299.50	1.15
Italian	155, 122, 115, 91, 79, 55, 40, 22, 16, 11, 0, 0;	12	10.17	156.46	1.13
Polish	93, 87, 52, 39, 36, 29, 24, 19, 8, 2, 2, 0;	12	9.00	94.93	1.23
Romanian	165, 104, 78, 75, 64, 60, 18, 0, 0, 0, 0, 0;	12	7.63	169.40	1.13
Russian	473,471,371,317,216,116,109,88,49,30,22,16;	12	10.55	457.54	1.18
Slovak	473,461,315,275,181,104,71,58,43,19,19,7;	12	11.15	467.22	1.06
Spanish	141, 102, 71, 51, 44, 41, 21, 6, 6, 3, 0, 0;	12	7.94	143.67	1.30
Ukrainian	310,282,193, 175, 118, 114, 50, 36, 23, 10, 4, 0;	12	10.14	310.51	1.23

The mean is $\overline{c}$ = 1.18, s_c = 0.0670, and the intervals are $c \in$< 1.0487, 1.3113>, $\overline{c} \in$<1.1404, 1.2196>.

3. Comparative treatment

First of all, we must state that the set of phenomena presented above does not allow us to set up founded hypotheses; in the best case we can make some conjectures. Neither the number of languages nor the diversity of language phenomena is sufficient. Diversification is often taken into account in grammar without regard to the frequency of phenomena. Grammarians content themselves with listing the existing cases; in language didactics frequency plays a more important role, but it is used only implicitly. Looking at Table 11, where the mean coefficients *c* are presented, one could conjecture that $\overline{c}$ increases from class building encompassing a complete field of phenomena to diversification of individual entities. The strongest diversification is that of meaning of independent full words followed by that of modifying affixes. The fact that in some languages affixes do not exist does not change anything. Affixes can diversify only if they exist.

Table 11 shows that if one takes into account class diversification, i.e. a classification comprising a complete inventory such as that of letters, word classes (parts of speech), a lexical word class or possible rhythmic units, the coefficient c is small. Phonic phenomena are on the lower end of the scale, semantic phenomena at the upper end. Morphological phenomena are somewhere in the middle. If one performed a different classification of parts of speech, e.g. with a stronger emphasis on syntax, it could be expected that the coefficient *c* would increase. On the other hand, diversifications taking into account whole classes (e.g. inventories) yield smaller c than diversification of individual entities (e.g. word meanings).

The end of this examination is open. We did not take into account any boundary conditions and only a small number of (available) phenomena. Nevertheless, two facts can be observed: crosslinguistic constancy of *c* for the same phenomenon and different magnitudes for different phenomena. A scaling of linguistic phenomena would be premature.

Table 11
Comparison of mean $\overline{c}$

Category	$\overline{c}$	s_c	**Int. *c***	**Int.** $\overline{c}$
1. Sounds, phonemes, letters	1.05	0.02	<1.00, 1.10>	<1.04, 1.06>
2. Word classes (parts of speech)	1.14	0.08	<0.98, 1.29>	<1.08, 1.18>
3. Rhythmic patterns	1.14	0.11	<0.92, 1.36>	<1.10, 1.18>
4. Paradigmatic classes	1.15	0.05	<1.05, 1.26>	<1.10, 1.20>
5. Colour classes	1.18	0.07	<1.05, 1.31>	<1.14, 1.22>
6. Prepositions, postposition, conjunctions	1.24	0.11	<1.03, 1.46>	<1.17, 1.32>
7. Case diversification	1.33	-	-	-
8. Allomorphs of plural	1.35	0.20	<0.97, 1.74>	<1.29, 1.41>
9. Affixes (Meaning diversification)	1.39	0.17	<1.06, 1.71>	<1.33, 1,45>
10. Words (Meaning diversification)	1.47	0.21	<1.06, 1.88>	<1.44, 1.50>

The computation of values in Table 11 has been performed to 4 decimal places and rounded to 2 places. A graphic presentation of the means and their intervals (second and last column of Table 11) can be seen in Figure 1.

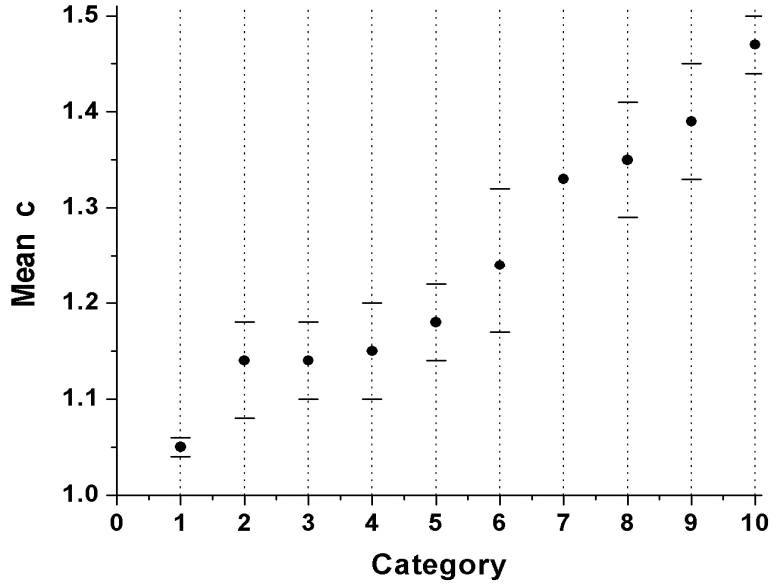


Figure 1. The positioning of linguistic categories regarding mean c

References

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 648-659*. Berlin/New York: de Gruyter.

Altmann,G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8, 130-139.*

Beöthy, E., **Altmann, G.** (1984a). Semantic diversification of Hungarian verbal prefixes. III. "föl-", "el-", "be-". *Glottometrika 7, 45-56.*

Beöthy, E., **Altmann, G.** (1984b). The diversification of meaning of Hungarian verbal prefixes. II. ki-. *Finnisch-Ugrische Mitteilungen 8, 29-37.*

Best, K.-H. (1991). *Von*: Zur Diversifikation einer Parttikel des deustchen. In: Rothe (1991): *94-104.*

Best, K.-H. (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1(2), 144-147.*

Best, K.-H. (2004/2005). Laut- und Phonemhäufigkeit im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Best, K.-H. (2008). Zur Diversifikation deutscher Hexameter (submitted).

Best, K.-H. (2008). Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 45-53.*

Brüers, N., **Heeren, A.** (2004). Pluralallomorphe in Briefen Heinrich von Kleists. *Glottometrics 7, 85-90.*

Dietze, J. (1982). Grapheme und Graphemkombinationen der russischen Fachsprache. *Glottometrika 4, 80-94.*

Drobisch, M.V. (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Lepzig. Philologisch-historische Classe 18, 73-139.*

Drobisch, M.V. (1968a). Weitere Untersuchungen über die Formen des Hexameters der Vergil, Horaz und Homer. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Philologisch-historische Classe 20, 16-53.*

Drobisch, M.V. (1968b). Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Lepzig. Philologisch-historische Klasse 20, 138-160.*

Drobisch, M.W. (1872). Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, 24, 1-33.*

Drobisch, M.W. (1875). Ueber die Gesetzmässigkeit in Goethe's und Schiller's Distichen. In *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 27, 8-34-146.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 82-89.*

Fry, D.B. (1947). The frequency of occurrence of speech sounds in Southern English. *Archives néerlandaises de phonétique expérimentale 20, 103-106.*

Fuchs, R. (1991). Semantische Diversifikation der deutschen Präposition *auf*. In Rotthe (1991): *105-115*.

Grigor´ev, V.I. (1980a). O dinamike raspredelenija bukv v tekste. In: *Aktual´nye voprosy strukturnoj i prikladnoj lingvistiki. Sbornik statej: 40-48.* Moskva.

Grigor´ev, V.I. (1980b). Frequency distribution of letters and their ranks in a running text. In: *Symposium Computational Linguistics and Related Topics. Summaries: 43-47*. Tallinn.

Grzybek, P. Kelih, E. (2006). Towards a general model of grapheme frequencies for Slavic languages. In: Garabík, R. (Ed.), Computer Treatment of Slavic and East European Languages: 73-87.. Bratislava: Veda.

Grzybek, P., Kelih, E. (2003). Graphemhäufigkeiten (am Beispiel des Russischen. Teil I. Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie 31, 131-162.*

Hammerl, R., Sambor, J. (1991). Untersuchungen zur Verteilung der Bedeutungen der polyfunktionalen polnischen Präposition *w* im Text. In: Rothe (1991): *127-137*

Hennern, A. (1991). Zur semantischen Diversifikation von „in“ im Englischen. In: Rothe (1991): *116-126.*

Job, M. (1974). *Untersuchungen zur Frequenz der Phoneme im Georgischen*. Unveröffentlichte Seminararbeit, Ruhr-Universität Bochum.

Kalinina E.A. (1968). Izučenie leksiko-statističeskich zakonomernostej na osnove vjerojatnostnoj modeli. In: *Statistika reči 64-107*. Lenngrad.

Kaliuščenko, V.D. (1988). *Deutsche denominale Verben*. Tübingen: Narr**.**

Kuße, H. (1991). *A* und *no* in N.M. Karamzins Pis´ma Russkogo Putešestvennika. In: Rothe (1991): *173-182.*

Meier, H. (1964). *Deutsche Sprachstatistik*. Hildesheim: Olms.

Meuser, K., Schütte, J.M., Stremme, S. (2008). Pluralallomorphe in den Kurzgeschichten von Wolfdietrich Schnurre. *Glottometrics 17, 2008, 20-25.*

Nemcová, E. (1991). Semantic diversification of Slovak verbal prefixes. In: Rothe (1991): *67-74.*

Nemcová, E. (2007). Zur Diversifikation des Bedeutungsfeldes slowakischer verbaler Präfixe. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text:499-508.* Berlin/New York: Mouton de Gruyter.

Ol´chin, P. (1907). Pervaja opora pri postroenii racional´noj stenografii. *Stenograf 4-5, 114-118*

Pääkkönen, M. (1994). Graphemes and context: statistical data on the graphology of standard Finnish. *Glottometrika 14, 1-53.*

Pawlowski, A. (1999). The quantitative approach in cultural anthropology: Application of linguistic corpora in the analysis of basic colour terms. *Journal of Quantitative Linguistics 6(3), 222-234.*

Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Word frequency and arc length. *Glottometrics 17, 18-44.*

Popescu, I.-I., Mačutek J., Altmann, G. (2008). *New Aspects of Word Frequencies,* Chapter 5. Arc length (pending)

Proskurin, N. (1933). Podsčety častoty liter i komplektovka šrifta. In: *Revoljucija i pis´mennost´. Sbornik I: 72-82.* Moskva-Leningrad.

Rademacher, A. (1974). *Untersuchungen zu den Buchstabenhäufigkeiten des See-Dajakischen.* Unveröffentlichte Seminararbeit, Ruhr-Universität Bochum

Roos, U. *(1991). Diversifikation der japanischen Postposition „-ni“*. In: Rothe (1991): *75-82.*

Rothe, U. (1986). *Die Semantik des textuellen* et. Frankfurt: Lang.

Rothe, U. (1990). Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina. *Glottometrika 11, 111-121.*

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Rothe, U. (1991a). Diversification of the case in German: genitive. In Rothe (1991): *140-156.*

Rothe, U. (1991b). Diversification processes in grammar: an introduction. In: Rothe (1991): *3-32.*

Sambor, J. (1989). Polnische Version des Projekts “Sprachliche Synergetik. Teil I. Quantitative Lexikologie. *Glottometrika 10, 171-197.*

Schulze, E. (1974). *Untersuchungen zu den Buchstabenhäufigkeiten des Hawaiischen.* Unveröffentlichte Seminararbeit, Ruhr-Universität Bochum

Schweers, A, Zhu, J. (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe (1991): *157-165.*

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5(3), 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček: 294-312*. Trier: Wissenschaftlicher Verlag

Glottometrics 17, 2008, 109-114

History of Quantitative Linguistics

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

XXXV. Moritz Wilhelm Drobisch (1802-1896)

Geb. 16.8.1802 Leipzig, gest. 30.9.1896 Leipzig. Mathematiker und Philosoph. Nach dem Schulbesuch in Leipzig und Grimma Studium in Leipzig. 1823 Promotion; 1824 Habilitation. 1824-1826 Privatdozent, ab 1826 a.o. Prof. in Leipzig; noch im gleichen Jahr bis 1868 o. Professor für Mathematik Ab 1842 gleichzeitig Prof. für Philosophie in Leipzig. 1868 verzichtet er auf die Professur für Mathematik. 1840/41 Rektor der Universität, insgesamt siebenmal Dekan der Philosophischen Fakultät. Ab 1835 Mitglied der *Fürstlich Jablonowskischen Gesellschaft*. 1846 wurde die *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig* gegründet, woran Drobisch einen maßgeblichen Anteil hatte.

Drobischs Bedeutung für die Quantitative Linguistik beruht auf seinen Vers-Studien. Von einer intensiven Befassung mit schöner Literatur ist bereits für das Jahr 1823 und für die späteren Lebensphasen („zur Erholung“) mehrfach die Rede: „Musik und Litteratur begleiten Drobisch auf allen seinen Lebenswegen. Ihnen und der Bewegung in der freien Natur widmet er alle seine freien Stunden“ (Neubert-Drobisch 1902: 15f., 106, 118). 1826 besuchte Drobisch anlässlich eines Aufenthalts in Göttingen den Mathematiker Gauß. Dazu heißt es: Gauß „sagte hierauf, er möchte schon einmal mit Hermann über Metrik sprechen, die ihn [gemeint: Gauß, Verf.] sehr interessiere, es seien hier ohne Zweifel noch viele und glänzende mathematische Entdeckungen zu machen: er glaube, die Metriker hätten noch gar keine mathematisch bestimmten Begriffe über kurze und lange Silben, geschweige über Accent und mehr. Unsere Lehre vom Schall sei überhaupt noch sehr unvollkommen. So frage es sich, welches der mathematische Unterschied zwischen den Schallwellen sei, die ein a und denen, die ein c zu hören geben. Man habe hierüber bis jetzt nur physiologische Untersuchungen angestellt, unter denen er von Kempelers Werk über die menschliche Stimme besonders rühmte. Auf Webers Wellenlehre, auf die ich ihn aufmerksam machte, war er sehr begierig.“ (Neubert-Drobisch 1902: 23)

Die Vers-Studien entstanden erst ab 1866 und damit viele Jahre später neben anderen Themen als Abhandlungen für die *Königlich-Sächsische Gesellschaft der Wissenschaften.*

Drobischs Zielsetzung bei seinen Hexameter-Studien besteht darin, Gesetzmäßigkeiten der Verteilung unterschiedlicher Hexameterformen herauszufinden. Er stellt sich vor, dass dabei eine „durchschnittliche Gesetzmäßigkeit“ (Drobisch 1866: 75) für die in einer Sprache produzierten Werke zu entdecken sein wird, aber auch das spezifische Gepräge, das einzelnen Autoren, bestimmten Zeitphasen oder auch dem jeweiligen Zustand einer Sprache zugemessen werden kann.

Bei seiner ersten derartigen Untersuchung widmet er sich den Hexameterformen lateinischer Autoren. Dazu stellt er zunächst für die ersten vier Versfüße die 16 Formen von Hexametern auf, die nach Stellung und Zahl von Daktylen und Spondeen möglich sind; die letzten beiden Versfüße lässt er außer Betracht; „die letzten zwei [= Versfüße; Verf.] sind fixiert“ (Strauß u.a. 1984). Eine Hexameterform kann dann z.B. als „sdds“ bestimmt werden. Das bedeutet, dass die ersten vier Versfüße aus der Folge Spondeus – Daktylus – Daktylus –

Spondeus bestehen. Dann untersucht er 16 Texte bzw. Kompilationen von 15 Autoren – Vergil ist mit zwei Texten vertreten – daraufhin, wie häufig bei ihnen in Textabschnitten von jeweils 80 Versen jede Hexameterform vorkommt und addiert sodann diese Ergebnisse für insgesamt 560 Verse, wenn denn soviel Text zur Verfügung steht. Im Anschluss werden die tabellarisch erfassten Verteilungen daraufhin verglichen, wie ähnlich oder auch wie verschieden sie sind. Im Ergebnis stellt Drobisch fest, wie ein „mittlere[r] lateinische[r] Hexameter" (Drobisch (1866:.137) aussieht, dass in den früheren Hexametern der Spondaeus stärker überwiegt als in den späteren, außerdem wie das Verhältnis der Zäsurentypen sich ändert, und schließt: „Ich glaube im Vorstehenden die a r c h i t e k t o n i s c h e n G e s e t z e des m i t t l e r e n lateinischen Hexameters dargelegt ... zu haben (Drobisch 1866: 138). Solche Untersuchungen haben für ihn sowohl eine philologische als auch eine philosophische Bedeutung.

Eingangs dieser Untersuchung schneidet Drobisch noch ein anderes Thema an, das vor ihm bereits Förstemann (1852) behandelt hat: das Verhältnis zwischen Konsonanten und Vokalen. Er kritisiert Förstemann, weil er nicht sage, auf welcher Basis er seine Daten erhoben habe, und stellt Vergleichszahlen aus dem Lateinischen vor (Drobisch 1866: 75ff., Fußnote).

Mit den vielen Daten, die Drobisch in der genannten und mehreren weiteren Untersuchungen erhoben hat, und seinem Anspruch, Gesetzmäßigkeiten aufzudecken, gehört er zu den Pionieren der Quantitativen Linguistik im 19. Jahrhundert. Er sollte aber noch stärker rezipiert werden: Sein Name ist bekannt; eine angemessene Rolle spielt er bisher in der Quantitativen Linguistik aber noch nicht. So findet man im ganzen Handbuch der Quantitativen Linguistik nur eine einzige Erwähnung (Köhler 2005: 3); in Meiers Zeittafel fehlt er ganz (Meier 1967: 349). Wenigstens Herdan (1966: 206-209), Grotjahn (1979: 205ff.) und Altmann (1981, 1988: 40-42) nutzen Drobischs Ergebnisse für ihre Zwecke. Job (1981: 235) widmet drei seiner Studien kurze Kommentare.

Um Drobischs Untersuchungen für die Zwecke der gegenwärtigen Quantitativen Linguistik noch weiter fruchtbar zu machen, kann man z.B. die verschiedenen Formen der Hexameter als ein Diversifikationsphänomen auffassen (Altmann 1991, 2005: 652). Unter diesem Gesichtspunkt wurden sie bisher offenbar noch nicht betrachtet. Zwecks Demonstration wird hier die Zipf-Mandelbrot-Verteilung

$$P_x = \frac{(b+x)^{-a}}{F(n)}, \quad x = 1,2,3,..n, \qquad F(n) = \sum_{i=1}^{n} (b+i)^{-a}$$

an die Daten der ersten 560 Verszeilen des Versepos *Aeneis* von Vergil (Drobisch 1866: 81, 83) angepasst , vgl. Tabelle 1). Eine ausführlichere Behandlung seiner umfangreichen Daten findet man in Best (2008).

Tabelle 1
Anpassung der Zipf-Mandelbrot-Verteilung an die Hexameterformen in Vergils *Aeneis*

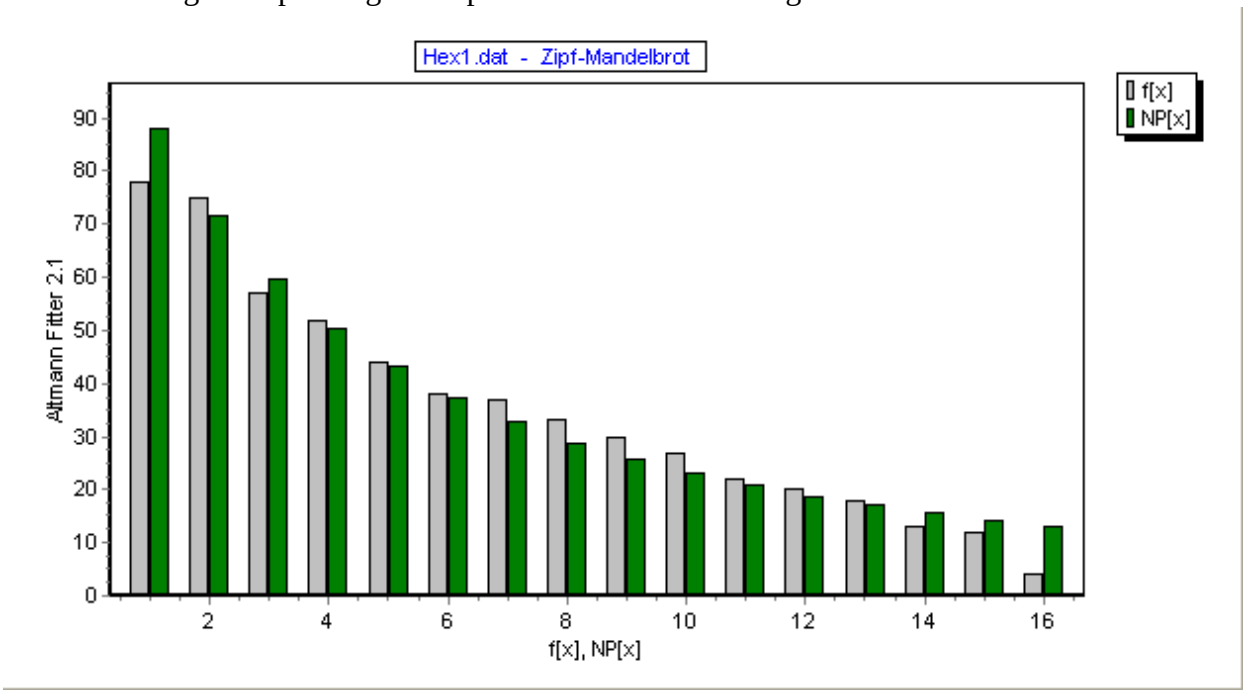
x	Hexameterform	n_x	NP_x
1	dsss	78	88.14
2	ddss	75	71.75
3	sdss	57	59.60
4	dsds	52	50.32

5	ddds	44	43.09
6	ssds	38	37.32
7	ssss	37	32.66
8	dssd	33	28.83
9	sdds	30	25.65
10	dsdd	27	22.97
11	ddsd	22	20.70
12	sdsd	20	18.75
13	sssd	18	17.07
14	ssdd	13	15.61
15	sddd	12	14.33
16	dddd	4	13.20
a = 1.9012	b = 7.7501 n = 16	FG = 12	X^2 = 11.587 P = 0.48

Legende zur Tabelle
d = Daktylus; s = Spondeus
x = Rang der Hexameterform
n_x = beobachtete Häufigkeit der Hexameterform
NP_x = aufgrund der Zipf-Mandelbrot-Verteilung berechnete Häufigkeit der Hexameterform
a, b, n = Parameter der Zipf-Mandelbrot-Verteilung
FG = Freiheitsgrade
X^2 = Chiquadrat
P = Überschreitungswahrscheinlichkeit des Chiquadrats

Das Ergebnis zeigt mit P = 0.48, dass die gewählte Verteilung sehr gut geeignet ist, als Modell für die beobachteten Werte verwendet zu werden. Es kommen noch eine ganze Reihe weiterer Modelle in Frage, die sowohl theoretisch als auch vom Ergebnis her verwendet werden könnten. Die folgende Graphik (Abb. 1) veranschaulicht das gute Ergebnis.

Abbildung 1. Anpassung der Zipf-Mandelbrot-Verteilung an die Daten von Drobisch



Mit diesem Beispiel sollte gezeigt werden, dass Drobischs Untersuchungen zu Versen weiterhin für die Ziele der gegenwärtigen Quantitativen Linguistik genutzt werden können; die Möglichkeiten dazu sind noch nicht ausgeschöpft. Bisher stand Drobischs erste Untersuchung von 1866 im Vordergrund des Interesses; einige seiner späteren Arbeiten sind offenbar hier zum ersten Mal bibliographisch erfasst. Vielleicht ergibt sich daraus ja ein Anstoß, sich noch einmal intensiver mit seinen Arbeiten zu befassen. Untersuchungen zur Diversifikation der deutschen und lateinischen Hexameter sind in Arbeit (Best 2008a, b).

Literatur

Altmann, Gabriel (1981). The homogenity of metric patterns in hexameter. In: Grotjahn, Rüdiger (ed.), *Hexameter Studies* (S. 137-150). Bochum: Brockmeyer.

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

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

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 646-658). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2008a). Zur Diversifikation deutscher Hexameter. In Arbeit.

Best, Karl-Heinz (2008b). Zur Diversifikation lateinischer Hexameter. In Arbeit.

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

Grotjahn, Rüdiger (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum: Brockmeyer.

Herdan, Gustav (1966). *The advanced theory of language as choice and chance*. Berlin/ Heidelberg/ New York: Springer.

Job, Ulrike (1981). Annotated bibliography on the statistical study of hexameter verse. In: Grotjahn, Rüdiger (ed.), *Hexameter Studies* (S. 226-262). Bochum: Brockmeyer.

Köhler, Reinhard (2005). Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 1-16). Berlin/ N.Y.: de Gruyter.

Meier, Helmut (1967). *Deutsche Sprachstatistik.* Hildesheim: Olms.

Strauss, U., Sappok, Ch., Diller, H.J., & Altmann, G. (1988). Zur Theorie der Klumpung von Textentitäten. In: Rothe, U. (ed.), *Glottometrika 7* (S. 73-100). Bochum: Brockmeyer.

Schriften von Moritz Wilhelm Drobisch (Monographien und Beiträge zur Metrik)

(Die Liste stützt sich hinsichtlich der Monographien weitgehend auf die Angaben von Neubert-Drobisch 1902, ergänzt vor allem um die für die Quantitative Linguistik einschlägigen Arbeiten zur Metrik. Es handelt sich also nicht um ein vollständiges Werkverzeichnis. Eine ganze Reihe weiterer Untersuchungen, vor allem naturwissenschaftlichen Inhalts, findet sich in den Berichten über die Verhandlungen der *Königlich-Sächsischen Gesellschaft zu Leipzig*, sowohl in der *Philologisch-Historischen Classe* als auch in der *Mathematisch-Physischen Classe.*)

1824. *Theoriae analyseos geometricae prolusio*. Diss. Leipzig. Leipzig: Glueck.
1825. *Grundzüge der ebenen und körperlichen Trigonometrie*. Leipzig: Baumgärtner.
1826. *De vera lunae figura observationibus determinanda disquisitio, annexa appendice de interiori terrae natura*. Leipzig: Carl Cnobloch.
1827. *Ad selenographiam methematicam symbolae*. Leipzig: Karl Phil. Melzer.
1832. *Philologie und Mathematik als Gegenstände des Gymnasialunterrichts betrachtet*. Leipzig: Carl Cnobloch.
1834. *Grundzüge der Lehre von den höheren numerischen Gleichungen*. Leipzig: Leopold Voß.
1834. *Beiträge zur Orientierung über Herbarts System der Philosophie*. Leipzig: Leopold Voß.
1836. *Neue Darstellung der Logik nach ihren einfachsten Verhältnissen mit Rücksicht auf Mathematik und Naturwissenschaft*. Leipzig: Leopold Voß (Bis 1887 insgesamt 5 Auflagen)
1837. *Quaestionum methamatico-psychologicarum fasciculus I.* Leipzig: Leopold Voß.
1840. *Grundlehren der Religionsphilosophie*. Leipzig: Leopold Voß.
1842. *Empirische Psychologie nach naturwissenschaftlicher Methode*. Leipzig: Leopold Voß. (²1898)
1846. *Über die mathematische Bestimmung der musikalischen Intervalle*. Leipzig: Leopold Voß.
1850. *Erste Grundlehren der mathematischen Psychologie*. Leipzig: Leopold Voß. (²1857)
1852. *Über musikalische Tonbestimmung und Temperatur*. Leipzig: Leopold Voß.
1864. *De philosophia scientiae naturali insita*. Leipzig: Leopold Voß.
1866. Ein statistischer Versuch über die Formen des lateinischen Hexameters. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 18, 75-139.*
1867. *Die moralische Statistik und die menschliche Willensfreiheit*. Leipzig: Leopold Voss.
1868. Weitere Untersuchungen über die Formen des Hexameter des Vergil, Horaz und Homer. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 20, 16-65.*
1868. Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 20, 138-160.*
1871. Über die Classification der Formen des Distichon. *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 23, 1-33.*
1872. Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 24, 1-33.*
1873. Ueber die Unterschiede in der Grundanlage des lateinischen und griechischen Hexameters. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 25, 7-32.*
1875. Ueber die Gesetzmässigkeit in Goethe's und Schiller's Distichen. In *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen 27, 8-34-146.*
1876. *Über die Fortbildung der Philosophie durch Herbarth*. Leipzig: Leopold Voss.
1885. *Kants Dinge an sich und sein Erfahrungsbegriff*. Leipzig: Leopold Voss.

Über Drobisch

ADB: Allgemeine Deutsche Biographie. 48. Band. Berlin: Duncker & Humblot 1971. (Neudruck der 1. Auflage von 1904.)

DBE: Deutsche Biographische Enzyklopädie (DBE). **Bd. 2.** Hrsg. v. Walter Killy. München u.a.: K.G. Sauer 1995.

NDB: Neue deutsche Biographie. 4. Bd. Hrsg. von der historischen Kommission bei der bayerischen Akademie der Wissenschaften. Berlin: Duncker & Humblot 1959.

Moritz Wilhelm Drobisch: http://www.uni-leipzig.de/~agintern/uni600/ng167d.pdf (mit Portrait).

Moritz Wilhelm Drobisch anlässlich seines 200. Geburtstages. Mit einem Vorwort von Uwe-Frithjof Haustein und Beiträgen von Gerald Wiemers und Lothar Kreiser. Verlag der Sächsischen Akademie der Wissenschaften zu Leipzig; In Kommission: Stuttgart/ Leipzig: Hirzel 2003. (Kleine Festschrift, die den „Anteil Drobischs an der Gründung der Königlich Sächsischen Gesellschaft der Wissenschaften 1846“ sowie „seine Beiträge zur Entwicklung der Logik, deren Niveau er nicht zuletzt durch sein Lehrbuch über Jahrzehnte bestimmte“ würdigt (Vorwort v. Haustein, 5).

Neubert-Drobisch, Walter (1902). *Moritz Wilhelm Drobisch: ein Gelehrtenleben*. Leipzig: Dieterich’sche Verlagsbuchhandlung Theodor Weicher.

Poggendorff, Johann Christian (1863). *Biographisch-literarisches Handwörterbuch zur Geschichte der exacten Wissenschaften. 1. Band. A – L.* Leipzig: Verlag von Johann Ambrosius Barth.

Karl-Heinz Best, Göttingen