Glottometrics 19
2009

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. 19 (2009), Lüdenscheid: RAM-Verlag, 2009. 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 19 (2009) **ISSN 2625-8226**

Contents

Glottometrics 19, 2009, 1-10

Wortlängen im Englischen

Karl-Heinz Best, Göttingen

Abstract. The aim of this paper is to show that word lengths in English texts follow certain distribution laws. The findings lend support to the theory of word length distributions (Wimmer et al. 1994, Wimmer & Altmann 1996) once more.

Keywords: Word length, English

0. Vorbemerkung

In diesem Beitrag werden Daten aus Wortlängenuntersuchungen von Elderton (1949) und Herdan (1960, 1966) darauf hin getestet, ob sie Gesetzmäßigkeiten folgen, die sich bisher im Göttinger *Projekt Quantitative Linguistik* in mehreren Arbeiten zum Englischen bewährt haben, oder ob ggfs. neue Modelle angewendet werden müssen.

1. Eldertons Pioniertat

Auf Elderton (1949)[1], der etwa gleichzeitig mit Čebanov (1947), aber noch vor Fucks (1955, 1956) zu den ersten Autoren gehört, die die Idee fassten, dass Wortlängen einem mathematischen Modell entsprechend verteilt sein sollten, macht Grzybek (2006:19ff.) nachdrücklich aufmerksam. Elderton (1949: 436) berichtet von J.B. Molony, der Wortlängen (gemessen nach der Zahl ihrer Silben) in einem englischen Text erhoben hatte und eine Ähnlichkeit dieser Daten mit der geometrischen Reihe vermutete (Daten zu diesem Text in Herdan 1960: 184). Elderton ging offenbar als erster diesem Hinweis nach und untersuchte englische Texte in etwas größerem Stil daraufhin, wie viele Wörter mit einer, zwei, drei etc. Silben in Texten, Textausschnitten oder auch Korpora vorkommen. Er verfolgte damit eine andere Hypothese als Čebanov und Fucks, die die Poisson-Verteilung als gutes Modell für Wortlängen betrachteten. Grzybek (2006: 23) zeigt am Beispiel einer der Dateien von Elderton, dass man die Idee, die geometrische Reihe sei möglicherweise für Wortlängen geeignet, zumindest in diesem Fall nicht verwerfen muss, macht aber auch darauf aufmerksam, dass sie nicht als generelles Modell für Wortlängenverteilungen in Frage komme. Offen geblieben ist bisher die Frage, ob es für Eldertons Daten überhaupt ein funktionierendes Modell gibt. Die folgende Untersuchung soll dafür eine Lösungsmöglichkeit aufzeigen. Dabei werden einige weitere Daten zu Wortlängen im Englischen berücksichtigt, die Herdan (1960, 1966) zusätzlich zu den Elderton übernommenen vorstellte. Auch Daten zu dem von Molony bearbeiteten Text werden berücksichtigt.

2. Modellierung der Wortlängen im Englischen

Als erster Schritt wurde Eldertons Idee nachgegangen, Wortlängen könnten der geometri-

[1] Zu Elderton siehe Best 2009.

schen Verteilung folgen. Er bemisst die Wortlänge meist nach der Zahl der Silben je Wort, in zwei Fällen auch nach der Zahl der Buchstaben. Elderton kommt jedoch selbst zu dem Schluss, dass seine Daten mit der geometrischen Verteilung nicht angemessen dargestellt werden können (Elderton 1949: 442). Eine Überprüfung der Dateien mit dem Chiquadrat-Test bestätigten sein Resumee.

In einem zweiten Schritt wurden die Ergebnisse herangezogen, die im Göttinger *Projekt Quantitative Linguistik* mit englischen Texten bereits erzielt wurden. Insgesamt gesehen haben sich dabei interessante Perspektiven ergeben. Während sich im Altenglischen noch die Hyperpoisson-Verteilung als ein gutes Modell sowohl für die Wortlängen in Gedichten als auch in Prosatexten erweist, findet mit neuer Aussprache im Mittelenglischen ein Übergang zu anderen Modellen statt (Zauner 2003). Bei neueren Prosatexten kommt neben der Hirata-Poisson-Verteilung (z.B. bei Spams: Jahn & Uckel 2008) und der gemischten Poisson-Verteilung (z.B. bei Pressetexten: Riedemann 1996) vor allem die positive Singh-Poisson-Verteilung in Betracht (z.B. bei Briefen: Frischen 1996, Zuse 1996). Insgesamt gesehen also ein recht heterogenes Bild. Dieser Eindruck wird noch dadurch verstärkt, dass Versuche mit Gedichten (Andrew Wilson, unveröffentlicht) auf die Poisson-Verteilung und die positive Poisson-Verteilung als mögliche Modelle hinweisen.

Bei der Bearbeitung der Daten, die Elderton (1949; teilweise auch in Herdan 1966: 285ff.) präsentiert, hat sich nun gezeigt, dass die Gedichte mit der Poisson-Verteilung und die übrigen Texte mit der positiven Singh-Poisson-Verteilung modelliert werden können. Die bisher beobachteten Haupttrends, denen die Wortlängen in Texten unterliegen, werden also bestätigt. Das Ergebnis ist auch deshalb beachtenswert, weil Elderton z.T. große Mischdateien gebildet hat, bei denen man mangels hinreichender Homogenität mit Problemen rechnen muss (Altmann 1992). Dennoch können für alle Textdateien die genannten Verteilungen genutzt werden.

3. Zu den verwendeten Modellen

Die Untersuchung steht wieder unter der Hypothese, dass Wortlängen in Texten sich gesetzmäßig und nicht etwa chaotisch verhalten. Die Ergebnisse, die in den folgenden Tabellen mitgeteilt werden, stützen diese Hypothese.

Bei den Gedichten wird die 1-verschobene Poisson-Verteilung

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

angepasst, bei allen übrigen Texten die positive Singh-Poisson-Verteilung

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

Die Verteilungen sind in dieser Form für x = 1, 2,… definiert, weil die Dateien mit x = 1 für

einsilbige Wörter beginnen. Beide Verteilungen ergeben sich aus der Theorie der Wortlängenverteilungen, wie sie von Wimmer u.a. (1994) sowie von Wimmer & Altmann (1996) entwickelt wurde.

4. Anpassung der Modelle an die Daten

Die Anpassungen wurden mit dem *Altmann-Fitter* (1997) durchgeführt und erbrachten folgende Ergebnisse:

Tabelle 1
Anpassung der Poisson-Verteilung an die Wortlängen in Gedichten[2] von Gray
(Elderton 1949: 438)

	Elegy		Sonnet on West		Eton	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	713	733.44	75	76.23	392	388.49
2	226	208.05	31	25.85	130	135.06
3	26	29.51	1	4.92	24	23.48
4	8	2.79			4	2.97
5	1	0.21				
	a = 0.2837 FG = 1 X^2 = 2.308 P = 0.13		a = 0.3390 FG = 1 X^2 = 4.173 P = 0.04		a = 0.3476 FG = 2 X^2 = 0.587 P = 0.75	

Bei *Sonnet on West* erhält man nur eine recht schwache Anpassung, was sicher auch auf den geringen Textumfang zurückzuführen ist.

Legende zu den Tabellen:
a, b: Parameter der Verteilung
n_x: beobachtete Häufigkeit des jeweiligen Fugenelements
NP_x: durch Anpassung der gewählten Verteilung berechnete Häufigkeit des Fugenelements
FG: Freiheitsgrade
X^2: Chiquadrat
P: Überschreitungswahrscheinlichkeit des Chiquadrats
C: Diskrepanzkoeffizient (Kontingenzkoeffizient) $C = X^2/N$ (N = Summe aller Beobachtungen), der statt P verwendet wird, wenn es sich um eine große Datei handelt oder $FG = 0$ ist.

Die Anpassung des gewählten Modells an die beobachteten Daten wird als erfolgreich angesehen, wenn $P \geq 0.05$ oder $C \leq 0.01$. Mit $0.01 > P \geq 0.05$ hat man ein Ergebnis, das nicht zufrieden stellt, aber auch nicht so schlecht ist, dass man es ganz verwerfen müsste.

[2] Elderton hat die Gedichte unterschiedlich bearbeitet; hier werden als Datengrundlage nur die Bearbeitungen der Texte als Gedichte berücksichtigt, nicht die Prosa-Bearbeitungen.

Tabelle 2
Anpassung der Poisson-Verteilung an die Wortlängen in Gedichten von Gray
(Elderton 1949: 438)

	Adversity		*Spring*		*Favorite Cat*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	205	207.93	206	204.69	178	179.76
2	81	74.83	61	63.40	56	51.95
3	7	13.47	6	9.82	6	8.29
4	5	1.77	6	1.10		
	a = 0.3599 FG = 1 X^2 = 1.238 P = 0.27		a = 0.3097 FG = 1 X^2 = 0.207 P = 0.65		a = 0.2890 FG = 1 X^2 = 0.963 P = 0.33	

Tabelle 3
Anpassung der Poisson-Verteilung an die Wortlängen
aller Gedichte Grays (Elderton 1949: 438)

x	n_x	NP_x
1	1769	1809.37
2	585	546.96
3	70	82.67
4	23	8.33
5	1	0.67
a = 0.3023 C = 0.0015		

Tabelle 4
Anpassung der positiven Singh-Poisson-Verteilung an die
Wortlängen in Briefen von Gray (Elderton 1949: 439)

x	n_x	NP_x
1	3987	3971.32
2	831	834.22
3	281	316.38
4	121	89.99
5	15	20.48
6	2	4.62
a = 1.1377 α = 0.5217 C = 0.0034		

Tabelle 5
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Textabschnitten von Carlyle (Elderton 1949: 439)

	French Revolution		*Past and Present*		*Heroes and Hero Worship*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	713	713.00	692	687.31	811	806.52
2	180	182.82	193	189.18	208	212.68
3	92	87.76	73	88.36	88	92.08
4	32	31.60	34	30.95	39	29.90
5	8	9.10	12	8.68	5	9.82
6	2	2.73	3	2.52		
	a = 1.4401 α = 0.5530 FG = 3 X^2 = 0.582 P = 0.90		a = 1.4012 α = 0.5856 FG = 3 X^2 = 4.448 P = 0.22		a = 1.2988 α = 0.5838 FG = 2 X^2 = 5.444 P = 0.07	

Tabelle 6
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in dem Mischtext aus Tabelle 5 (Elderton 1949: 439)

x	n_x	NP_x
1	2216	2211.97
2	581	582.25
3	253	266.77
4	105	91.67
5	25	25.20
6	5	7.14
a = 1.3745 α = 0.5715 FG = 3 X^2 = 3.300 P = 0.35		

Tabelle 7
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Macaulays Essays (Elderton 1949: 439)

	Clive		*Hustings and Milton*		*Mischdatei aus beiden*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	1260	1260.43	723	722.27	1983	1983.33
2	375	383.30	189	187.89	564	570.31
3	201	185.49	95	100.77	296	286.52
4	67	67.32	42	40.53	109	107.96
5	14	19.55	17	13.04	31	32.54
6	5	5.92	2	3.50|	7	8.18
7			1	1.00|	1	2.16
	a = 1.4518 α = 0.6190 FG = 3 X^2 = 3.196 P = 0.36		a = 1.6090 α = 0.5428 FG = 3 X^2 = 2.090 P = 0.55		a = 1.5072 α = 0.5899 FG = 4 X^2 = 1.256 P = 0.87	

Tabelle 8
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in 3 Textausschnitten von 2 Autoren (Elderton 1949: 440)

	Johnson		*Gibbon*		*Gibbon*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	1268	1263.68	1243	1242.07	617	615.07
2	423	419.36	408	420.44	206	213.48
3	195	208.61	224	217.84	118	110.65
4	77	77.83	103	84.65|	50	43.01
5	29	23.23	21	26.32|	6	13.38
6	8	7.28	1	8.69|	3	4.42
	a = 1.4924 α = 0.6492 FG = 3 X^2 = 2.447 P = 0.48		a = 1.5544 α = 0.6495 FG = 1 X^2 = 0.782 P = 0.38		a = 1.5549 α = 0.6595 FG = 3 X^2 = 1.256 P = 0.74	

Im Fall von Gibbon handelt es sich um 2 unterschiedlich zusammengestellte Textausschnitte aus *Decline and Fall.*

Tabelle 9
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Bibeltexten aus *Genesis* und *Exodus* (Elderton 1949: 440)

	alle Wörter		*ohne Namen*	
x	n_x	NP_x	n_x	NP_x
1	2943	2942.10	2939	2937.39
2	579	574.46	520	513.61
3	77	85.88	61	73.40
4	12	9.63|	11	7.87|
5	2	0.93|	2	0.73|
	a = 0.4485, α = 0.8948, FG = 1, X^2 = 2.074, P = 0.15		a = 0.4288 α = 0.8467 C = 0.0013	

Tabelle 10
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Shakespeares *Heinrich IV* (Elderton 1949: 441)

	Prosa-Passagen		Vers-Passagen	
x	n_x	NP_x	n_x	NP_x
1	10965	10943.37	9076	9065.08
2	2177	2152.95	1918	1931.55
3	430	500.63	476	480.48
4	99	87.31	108	89.64
5	23	12.18	4	15.24
6	2	1.42|		
7	2	0.15|		
	a = 0.6976 α = 0.6517 C = 0.0018		a = 0.7463 α = 0.6643 C = 0.0011	

Tabelle 11
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Essays von Bacon (Elderton 1949: 441)

x	n_x	NP_x
1	4640	4628.55
2	1080	1079.09
3	420	454.88
4	167	143.81
5	41	36.37
6	3	7.67
7	1	1.64
$a = 1.2646$ $\alpha = 0.5400$ $C = 0.0016$		

Damit kann festgestellt werden, dass an alle Dateien, die Elderton mitteilt, eine Verteilung angepasst werden kann, die aus der Theorie der Wortlängenverteilung hergeleitet werden kann. Für Johnson und Gibbon nennt Elderton (1949: 443-445) auch noch die Verteilung der Wortlängen, die in diesem Fall nach der Zahl der Konsonanten und Vokale und dann nach der Zahl der Buchstaben insgesamt bestimmt wurde. Versuche, an die Dateien, die alle Buchstaben berücksichtigen, ein Modell anzupassen, sind misslungen.

Auch bei der Datei, die womöglich am Anfang der Wortlängenuntersuchungen zum Englischen stand, lässt sich das gleiche Modell anpassen. Die Daten werden von Herdan (1960: 184) mitgeteilt, mit Verweis auf Elderton (1949: 136), wo zwar über den Versuch Molonys in den 1940er Jahren berichtet wird, aber keine entsprechenden Daten zu finden sind.

Tabelle 12
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen
in Fitzgeralds Übersetzung von Omar Khayyám, *Rubáiyát* (Herdan 1960: 184)

x	n_x	NP_x
1	1832	1828.02
2	420	420.13
3	88	91.65
4	12	14.99
5	5	2.21
$a = 0.6544$ $\alpha = 0.7692$ $FG = 2$ $X^2 = 4.322$ $P = 0.12$		

Testet man für diese Datei die geometrische Verteilung, so erhält man sowohl mit $P = 0.31$ als auch mit $C = 0.0015$ gute Ergebnisse; Molony, der nach Elderton (1949: 136) eine Ähnlichkeit seiner Daten mit der geometrischen Verteilung sah, hatte damit also für diesen Fall durchaus recht.

Drei weitere Dateien teilt Herdan (1966: 286) mit, die nicht schon bei Elderton zu finden sind. Auch in diesen Fällen kann die positive Singh-Poisson-Verteilung mit guten Ergebnissen angepasst werden.

Tabelle 13
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen in englischen Texten (Herdan 1966: 286)

	G.B. Shaw, *Dramatic Criticism*		Dewey A		Dewey B	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	584	581.91	66593	66559.11	72031	72009.23
2	184	185.92	9309	9494.37	12163	12148.69
3	72	75.87	3155	2820.23	2589	2687.70
4	29	23.22	476	628.29	516	445.96
5+	5	7.08	100	130.99	59	66.42
	a = 1.2242 α = 0.6817 FG = 2 X^2 = 2.276 P = 0.32		a = 0.8911 α = 0.4318 C = 0.0011		a = 0.6637 α = 0.5948 C = 0.0002	

Zu Dewey: Es handelt sich bei Dewey um einen Mischtext des Neuenglischen aus 100000 Wörtern. Die beiden Dateien unterscheiden sich dadurch, dass in Dewey A jede Variante einer Wortwurzel als eigenständiges Wort gewertet wurde, während in Dewey B die unterschiedlichen Wurzeln desselben Wortes als Vorkommen ein und desselben Wortes aufgefasst wurden. In beiden Fällen wurden nur diejenigen Klassen berücksichtigt, die mindestens zehnmal vorkamen (Herdan 1966: 286).

Auch wenn man sich nur auf die Wortlängen von Substantiven bezieht, kann das gleiche Modell angepasst werden (vgl Tab. 14).

Tabelle 14
Anpassung der positiven Singh-Poisson-Verteilung an die Wortlängen von Substantiven in Bunyans *Pilgrim's Progress* (Herdan 1966: 294)

	Teil I		Teil II		beide Teile zusammen	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	2429	2428.11	2451	2446.82	4880	4876.70
2	1129	1130.88	1105	1126.47	2234	2259.80
3	372	373.58	383	345.82	755	717.27
4	101	92.56	73	79.63	174	170.75
5	16	21.86	4	17.25	20	38.47
	a = 0.9910 α = 0.9639 FG = 2 X^2 = 2.353 P = 0.31		a = 0.9210 α = 0.9998 C = 0.0038		a = 0.9522 α = 0.9838 C = 0.0014	

Die vereinigte Tabelle ist nicht Herdan (1966) entnommen, sondern neu erstellt worden.

5. Ergebnis

Als Ergebnis dieser Untersuchung kann damit festgestellt werden, dass an alle Dateien, die Elderton und Herdan vorstellen, eines der bereits vielfach bewährten Modelle angepasst werden kann. Die geometrische Verteilung ist nur ausnahmsweise anwendbar. Die Ergebnisse,

die in mehreren Untersuchungen im Göttinger *Projekt Quantitative Linguistik* erzielt wurden, finden eine deutliche Bestätigung.

Ein eigenes Problem stellen die Untersuchungen dar, bei denen Wortlängen nicht nach Silben, sondern nach kleineren Einheiten (Buchstaben, Phoneme) bestimmt wurden. Für die beiden von Elderton (1949: 444f.) und die von Herdan (1966: 301) genannten Verhältnisse sind geeignete Verteilungen ungleich schwerer zu ermitteln. Der Grund dafür ist vermutlich darin zu sehen, dass zwischen Wort und Buchstabe/ Laut/ Phonem eine Ebene der Morpheme/ Silben anzusetzen ist, die sich bei der Modellierung als Störfaktor bemerkbar machen kann. Eine genauere Untersuchung dieser Wortlängenverhältnisse steht noch aus.

Literatur

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

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

Best, Karl-Heinz (2009). William Palin Elderton (1877-1962). Eingereicht.

Čebanov, Sergej Grigor'evič (1947). O podčinenii rečevych ukladov 'indoevropejskoj' gruppy zakonu Puassona. *Doklady Akademii Nauk SSSR. Tom 55/2, 103-106.*

Elderton, William P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society, Series A (General), Vol. CXII, Part IV, 436-445.*

Frischen, Jutta (1996). Word Length Analysis of Jane Austen's Letters. *Journal of Quantitative Linguistics 3, 80-84.*

Fucks, Wilhelm (1955). Theorie der Wortbildung. *Mathematisch-Physikalische Semesterberichte. Bd. 4, 195-212.*

Fucks, Wilhelm (1956). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Forschungsberichte 3, 7-21.*

Grzybek, Peter (2006). History and Methodology of Word Length Studies. The State of the Art. In: Grzybek, Peter (ed.). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues* (S. 15-90). Dordrecht: Springer.

Herdan, Gustav (1960). *Type-Token Mathematics. A Textbook of Mathematical Linguistics*. 's-Gravenhage: Mouton.

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

Jahn, Thomas, & Uckel, Annika (2008). Verteilung von Wortlängen in englischen Spam-E-Mails. *Glottometrics 17, 2008, 1-7.*

Köhler, Reinhard (1995). *Bibliography of quantitative linguistics*. Amsterdam: John Benjamins.

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

Riedemann, Hagen (1996). Word Length Distribution in English Press Texts. *Journal of Quantitative Linguistics 3, 265-271.*

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

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

Zauner, Thomas (2003). *Die Entwicklung der Wortlänge in der Sprachgeschichte des Englischen.* Göttingen, Magisterarbeit.

Zuse, Maria (1996). The Distribution of Word Length in English Letters of Sir Philip Sidney. *Journal of Quantitative Linguistics 3, 272-276.*

Software

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

Glottometrics 19, 2009, 11-42

Semantische Kombinierbarkeit von Komponenten in der Struktur der deutschen Komposita

Viktor Lewizkij[1], Yuliya Matskulyak, Cernivcy

Abstract. The semantics of nouns, adjectives and verbs as components of compound words was studied; also the peculiarities of their combination in the models N+N, A+N, V+N were analyzed. The main characteristics of the combinatory were established as combinatory range, strength (intensity) of a tie, and the number of significant ties.

Keywords: lexical-semantic subclass, compound noun, quantitative analysis, semantic identity/congruence.

1. Ziel

Die Problematik der zusammengesetzten Wörter in der deutschen Sprache stand immer im Zentrum des Interesses von Germanisten. Außer den bekannten grundlegenden Forschungen von J. Erben (Erben 2000), W. Fleischer (Fleischer 1969), W. Fleischer und I. Barz (Fleischer/Barz 1995), W. Henzen (Henzen 1957), F. Kluge (Kluge 1913), I. Kühnhold und H. Wellmann (Kühnhold/Wellmann 1973), I. Kühnhold, O. Putzer und H. Wellmann (Kühnhold/Putzer/ Wellmann 1978), W. Motsch (Motsch 1999), B. Naumann (Naumann 2000) und anderen, in denen die deutsche Wortbildung einschließlich der Wortzusammensetzung betrachtet wird, gibt es eine große Zahl von Forschungen, die den aktuellen Fragen der Wortbildung (Altmann, 1989) sowie der Struktur und der Klassifikation von Komposita (Fleischer/Barz 1995, Stepanova 1953, Barz 2000, Donalies 2002, Erben 2000, Naumann 2000), dem Funktionieren der Fugenelemente in Zusammensetzungen (Dressler/Barbaresi 1991, Gallmann 1999, Nübling 2004, Lindner 1998), den pragmatischen und syntagmatischen Beziehungen in zusammengesetzten Konstruktionen (DW 1991), dem Gebrauch von Komposita in verschiedenen Funktionalstilen (Levkovskaja 1960, Sajenko 2002, Chang 2005, Rings 2001, Feine 1997) und anderen Bereichen gewidmet sind.

Unter der großen Zahl von Arbeiten zur Kompositabildung im Deutschen sind nur wenige, in denen nicht die Struktur, sondern ihre Semantik erforscht wird (siehe, z.B., Schröder 1979, Steiner 2002, Vandermeeren 1999, Rothe 1988, Langer 1998).

Obwohl die Meinung vorherrscht, dass jedes Substantiv im Deutschen mit einem anderen beliebigen Substantiv ein neues zusammengesetztes Wort bilden kann (siehe, z.B., Langer 1998), sind die Verbindungsfähigkeiten der Substantive sowie anderer Wortarten in zusammengesetzten Wörtern nach wie vor wenig erforscht. Ist die Freiheit der Verbindung von Substantiven in einem Kompositum wirklich unbeschränkt? Wenn es doch irgendwelche Beschränkungen gibt: Wie und in welchem Maße offenbaren sie sich? Können wir das Ausmaß (die Breite) der lexikalischen Kombinierbarkeit von Kompositakomponenten untersuchen oder sogar berechnen? Können wir die Intensität des Zusammenhangs zwischen den Kompositakomponenten entdecken und ausmessen? Gibt es irgendwelche Gesetzmäßigkeiten bei der Zusammensetzung einzelner Wörter oder Wortklassen? Diese und andere Fragen sol-

[1] Address corrspondence to: Viktor Lewizkij, Radiščev-Str. 6/5, UA-58000. E-mail: lessja@gmail.com, Yuliya Matskulyak, Tschervonoarmijska-Str. 69/131, UA-58029. E-mail: juliamatskuliak@yahoo.de.

len beantwortet werden. So ist das Ziel unserer Arbeit die Untersuchung der semantischen Kombinierbarkeit von Komponenten in deutschen Substantivkomposita.

2. Materialien

Besonders geeignete Materialien für die Antwort auf die obengestellten Fragen bietet zweifellos die Belletristik, denn es kann davon ausgegangen werden, dass viele im Wörterbuch nicht fixierten Autorenkomposita-Neologismen in literarischen Texten vorkommen. Dabei muss auch berücksichtigt werden, dass die Wortzusammensetzung bestimmte Besonderheiten abhängig vom Funktionalstil haben kann. Darum soll die Stichprobe so organisiert werden, dass man dabei die Belletristik mit anderen Funktionalstilen vergleichen könnte. Von besonders großem Interesse sind unter Berücksichtigung der in der Forschung gestellten Aufgaben der publizistische und der wissenschaftliche Stil. Dementsprechend besteht die Stichprobe aus drei Textgruppen, die zur Belletristik (mehr als 340 000 Textwörter), zum Stil der wissenschaftlichen Literatur (mehr als 37 000 Wörter) und zur Publizistik (mehr als 32 500) gehören.

Aus diesen Texten wurden alle Substantivkomposita ausgewählt, insgesamt 17 179: 10 825 aus der Belletristik, 3 327 aus der wissenschaftlichen Literatur und 3 027 aus der Publizistik. Das Verhältnis der in der Stichprobe erhaltenen Komposita beträgt 3 : 1 : 1. Die Gesamtheit der belletristischen Texte bilden 16 Romane von 8 bekannten deutschen Autoren (H. Böll, G. Grass, St. Heym, H. Kant, S. Lenz, E. Loest, M. Walser, Ch. Wolf). In den Werken dieser Autoren wurde jede fünfte Seite analysiert. Die Gesamtheit der wissenschaftlichen Texte bilden 7 Monographien auf dem Gebiet der Biologie, Ökologie, Wirtschaft, Rechtswissenschaft, Soziologie, Philologie und Chemie. Die Untersuchung zur Publizistik wurde aufgrund der deutschen Zeitungen „Die Welt“ (1995), „Die Zeit“ (1988) und der Zeitschrift „Der Spiegel“ (1994) durchgeführt: Es wurden die Leitartikel auf der ersten Seite der Zeitungen und in den Zeitschriftenrubriken *Titel*, *Deutschland*, *Gesellschaft*, *Ausland* analysiert.

Der relative Fehler der Stichprobe δ für die belletristischen Texte beträgt 0,02, für die wissenschaftlichen 0,03 und für die publizistischen 0,04. Die Größe des relativen Fehlers der Stichprobe, die Repräsentation und die zeitliche Beschränkung der ausgewerteten Texte lassen also die Erwartung zu, dass eine hohe Zuverlässigkeit der zu erwartenden Ergebnisse gewährleistet ist.

3. Semantik der Unterklassen der zu erforschenden Substantive

Die Semantik von Konstituenten, aus denen ein Kompositum besteht, kann auf dem Niveau eines einzelnen Wortes oder einer Wortunterklasse untersucht werden. Da das Ziel unserer Forschung die Erschließung von allgemeinen Gesetzmäßigkeiten der semantischen Kombinierbarkeit von Kompositakomponenten ist, erarbeiten wir sie auf dem Niveau der Wortunterklasse.

Das Problem der semantischen Klassifikation eines Lexems, d.h. der objektiven Zerlegung des zu erforschenden Satzes von lexikalischen Einheiten nach bestimmten semantischen Unterklassen, ist bis heute ungelöst, obwohl in der Semasiologie mehrmals versucht wurde, eine Prozedur für die semantische Klassifikation mithilfe von Struktur-, psycholinguistischen und statistischen Methoden zu erarbeiten. In der Regel beruhen die meisten semantischen Klassifikationen auf einer intuitiven Analyse, obwohl selbstverständlich in den Forschungen, die der Analyse der Bedeutungsseite der Sprache gewidmet sind, eine ganze Reihe von Forde-

rungen zur Kategorisierung der lexikalischen Einheiten erarbeitet ist, die für die maximale Objektivität einer solchen Kategorisierung bestimmt sind.

Die semantischen Klassifikationen von Substantiven, Adjektiven und Verben in unserer Untersuchung wurden aufgrund aller Zusammensetzungen aus der Stichprobe, die nach dem Modell N + N (Substantiv + Substantiv), A + N (Adjektiv + Substantiv) und V + N (Verb + Substantiv) gebildet sind, erarbeitet. Dabei wurde die Verteilung nach der ersten Bedeutung im Wörterbuch DUDEN (1995) durchgeführt. Die vorgeschlagenen Klassifikationen scheinen für unsere Forschung am besten geeignet zu sein. Sie spiegeln alle semantischen Besonderheiten von Substantiven, Adjektiven und Verben, die als Bestimmungs- und Grundwörter auftreten, wider. Grundlegend war dabei die Verteilung nach allgemeinen Klassen. Es ist hinlänglich bekannt, dass die semantischen Charakteristika der Substantive sehr unterschiedlich sind (Smirnowa 1982); sie umfassen die Bezeichnungen von Lebewesen und Gegenständen, konkreten und abstrakten Begriffen, Eigen- und Gattungsnamen. Der Satz der semantischen Kategorien von Substantiven schließt folgende lexikalisch-semantischen Unterklassen ein (LSU):

1. Zur LSU „**Person**" zählen wir die Bezeichnungen von Menschen und außerirdischen Wesen, sowie Völkernamen, Verwandtschaftsbezeichnungen, Berufe, z.B.: *Mensch*, *Frau*, *Kind*, *Arbeiter*, *Mann*, *Gott*, *Führer* u.a.

2. Zur LSU „**Tiere**" gehören die Bezeichnungen von Tieren (Vögeln, Fischen und Insekten), z.B.: *Schwein*, *Pferd*, *Fisch*, *Hund*, *Löwe*, *Butt*, *Biene* u.a.

3. Zur LSU „**Somatismen**" werden Substantive gezählt, die Körperteile von Menschen und Tieren bezeichnen, z.B.: *Hand*, *Kopf*, *Blut*, *Ei*, *Leib*, *Brust* u.a.

4. Zur LSU „**Attribute des Menschen**": Kleider-, Schuhwerkbezeichnungen sowie ihre Teile, z.B.: *Schuh*, *Hose*, *Schmuck*, *Wäsche*, *Hemd*, *Jacke*, *Krone* u.a.

5. Zu den Einheiten der LSU „**Pflanzen**": Pflanzenbezeichnungen, ihre Bestandteile und Früchte, z.B.: *Kartoffel*, *Pflanze*, *Reis*, *Frucht*, *Apfel*, *Hafer*, *Pilz* u.a.

6. Zur LSU „**Stoffe und Materialien**", z.B.: *Wasser*, *Holz*, *Stein*, *Gas*, *Sand*, *Stoff*, *Beton* u.a.

7. Zur LSU „**Raum und Ort**": Substantive, die geographische und astronomische Objekte, die Natur und Orte bezeichnen, z.B.: *Land*, *Stadt*, *Welt*, *Straße*, *Ost*, *Wald*, *Lager* u.a.

8. Zur LSU „**Gebäude und Bauten**": räumliche Objekte, Bauten sowie einzelne Teile der Gebäude und Konstruktionen, z.B.: *Haus*, *Küche*, *Fenster*, *Kolonne*, *Büro*, *Kirche*, *Dach* u.a.

9. Zur LSU „**Gegenstände und Instrumente**": Substantive zur Bezeichnung von Einrichtungen, Geräten, darunter Verkehrsmittel, Waffen, Musikinstrumente, Möbel, Geschirr und ihre Bestandteile, z.B.: *Waffe*, *Buch*, *Auto*, *Panzer*, *Schiff*, *Maschine*, *Tasche* u.a.

10. Zur LSU „**Essen und Getränke**": Speisenamen, Getränke und ihre Bestandteile, z.B.: *Bier*, *Milch*, *Tee*, *Pfeffer*, *Fleisch*, *Brot*, *Schnaps* u.a.

11. Zur LSU „**Anzahl, Maßeinheiten**": Substantive zur Bezeichnung von Mengen und Portionen, Gesamtheit der Komponenten und Maßeinheiten, z.B.: *Atom*, *Teil*, *Kern*, *Mehrheit*, *Rest*, *Tropfen*, *Zusatz* u.a.

12. Zur LSU „**Bewegung**": Substantive, die Bewegung, Standortverlagerung, Lageveränderung bezeichnen, z.B.: *Reise*, *Verkehr*, *Flug*, *Übergang*, *Ausgang*, *Strömung*, *Ablauf* u.a.

13. Die LSU „**Tätigkeit, Aktion**" schließt Substantive ein, die Aktionen und menschliche Tätigkeiten, die mit körperlicher Einwirkung auf das Objekt verbunden sind, bezeichnen, z.B.: *Arbeit*, *Regierung*, *Bau*, *Dienst*, *Fertigung*, *Leistung*, *Trocknung* u.a.

14. Die LSU „**Dasein**" enthält Substantive, die die Existenz, ihren Anfang, ihr Ende, Ereignisse, Zustände, Zustands- oder Merkmalveränderungen bezeichnen, z.B.: *Leben*, *Tod*, *Frieden*, *Entwicklung*, *Schluss*, *Anfang*, *Bestand* u.a.

15. Zur LSU „**Possessorische Sphäre**" zählten wir Substantive, die das Besitzen, Erhalten und Verlieren bezeichnen, z.B.: *Gut, Besitz, Investition, Kapital, Einkauf, Einkommen, Erwerb* u.a.

16. Die LSU „**Mentale Sphäre**" umfasst Substantive, die mit der geistigen Tätigkeit des Menschen verbunden sind, z.B.: *Geist, Traum, Gedanke, Planung, Erinnerung, Entscheidung, Forschung* u.a.

17. Zur LSU „**Wahrnehmung**" gehören Substantive, die mit der Wahrnehmung (mithilfe der Rezeptionsorgane) verbunden sind, z.B.: *Sinn, Aussicht* u.a.

18. Die LSU „**Seelische Sphäre**" schließt die Bezeichnungen von Emotionen, Gefühlen und dem Willen ein, z.B.: *Wahl, Liebe, Trauer, Schrecken, Wunsch, Seele, Wille* u.a.

19. Die LSU „**Sprache und Rede**" enthält Substantive, die mit der sprachlichen Kommunikation, Informationsübermittlung und -beschaffung verbunden sind, z.B.: *Wörter, Sprache, Text, Brief, Zeitung, Kommando, Ausdruck* u.a.

20. Zur LSU „**Physiologische Sphäre**" gehören lexikalische Einheiten zur Bezeichnung physiologischer Wirkungen, Krankheiten, z.B.: *Hunger, Atem, Gesundheit, Pest, Schmerz, Gicht, Migräne* u.a.

21. Unter den Einheiten der LSU „**Verhalten und Handlungen**" sind Substantive zur Bezeichnung von Taten und Handlungen des Menschen, Zusammenarbeit und Beziehungen zu verstehen, z.B.: *Krieg, Kampf, Hilfe, Ehre, Handel, Auftrag, Reform* u.a.

22. Zur LSU „**Eigenschaften des Menschen**" zählten wir Substantive, die die Merkmale, Eigenschaften des Menschen bezeichnen oder seinen Charakter, sein Temperament angeben, z.B.: *Energie, Kraft, Unabhängigkeit* u.a.

23. LSU „**Naturerscheinungen und Zustände**" enthält Bezeichnungen von Erscheinungen und Zuständen der Umwelt, z.B.: *Natur, Dampf, Eis, Feuer, Regen, Schnee, Brand* u.a.

24. Zur LSU „**Physikalische Eigenschaften**" gehören Substantive zur Bezeichnung akustischer Phänomene, der Farben, des Lichts, Geschmacks, Geruchs, der Temperatur, z.B.: *Wärme, Licht, Ton, Temperatur, Schall, Akkord, Farbe* u.a.

25. Zur LSU „**Zeit, Alter**" zählen wir die Bezeichnungen von Zeitabschnitten, Zeiträumen, Alter, z.B.: *Zeit, Jahr, Nacht, Morgen, Winter, Jugend, Alter* u.a.

26. Die LSU „**Kennwerte und Eigenschaften der Gegenstände**" enthält Substantive zur Bezeichnung von Merkmalen und Kriterien z.B.: *Seite, Verfahren, System, Qualität, Wert, Rasse, Ordnung* u.a.

27. Zur LSU „**Veranstaltung, Spiel**" gehören Bezeichnungen von zu verschiedenen Zwecken organisierten Veranstaltungen, u.a. sportlichen, z.B.: *Markt, Sport, Programm, Konferenz, Streik, Skat, Revolution* u.a.

28. Die LSU „**Eigennamen**" enthält Namen von Menschen, historischen Persönlichkeiten, erdkundliche Namen, z.B.: *Europa, Bosnien, Hitler, Silvester, Marie, Weichsel* u.a.

29. Zur LSU „**Staat, seine Attribute**" gehören Substantive zur Bezeichnung der politischen Gesellschaftsorganisation und ihre Attribute, z.B.: *Staat, Wirtschaft, Recht, Militär, Reich, Verwaltung, Verfassung* u.a.

30. Unter Einheiten der LSU „**Dokumente, Geld**" – Bezeichnungen von schriftlichen Urkunden, offiziellen Zeugnissen und Geldscheinen, z.B.: *Steuer, Finanz, Geld, Preis, Lohn, Tarif, Akte* u.a.

31. Die LSU „**Termini**" enthält Wörter aus der beruflichen Tätigkeit, z.B.: *Kristallisation, Absorption, Extraktion, Adsorption, Diffusion, Priorität* u.a.

32. Zur LSU „**Sammelbezeichnungen von Menschen, Organisationen**" gehören Substantive zur Bezeichnung von in bestimmten Gruppen vereinigten Leuten, z.B.: *Volk, Partei, Schule, Betrieb, Familie, Bund, Amt* u.a.

33. Die LSU „**Abstrakte Begriffe**" umfasst z.B.: *Macht, Sicherheit, Druck, Glück, Inhalt, Begriff, Freiheit* u.a.

34. Die LSU „**Wissenschaft, Kultur, Traditionen**" enthält Substantive, die Wissenschaftszweige, Religionsströmungen, Kunstrichtungen sowie Zeichen, Symbole, Signale bezeichnen, z.B.: *Kultur, Kunst, Wissenschaft, Lehre, Alarm, Signal, Religion* u.a.

Bei den Adjektiven unterscheiden deutsche Grammatiken traditionell qualitative, possessive sowie quantitative. Die semantische Klassifikation von Adjektiven in unserer Arbeit schließt folgende Unterklassen ein:

1. Die LSU „**Größe, Abstand**" enthält Adjektive mit dimensionalen Angaben, z.B.: *groß, klein, hoch* u.a.

2. Die LSU „**Anzahl**" umfasst z.B.: *einzeln, doppelt, halb, gesamt* u.a.

3. Zur LSU „**Zeit**" rechnen wir Einheiten zu, die Zeit, Dauer und das Alter charakterisieren, z.B.: *neu, jung* u.a.

4. Die LSU „**Bewertung**" schließt überwiegend die subjektive Charakteristik des Gegenstandes bezüglich seines Wertes, seiner Bedeutung, Funktion ein, z.B.: *besonder…, speziell, edel, normal, wohl, original* u.a.

5. Zur LSU „**Richtung**" zählen Adjektive, die die Bewegungsrichtung und das Verhältnis zwischen Gegenständen im Raum charakterisieren, z.B.: *quer* u.a.

6. Die LSU „**Zustand**" enthält Adjektive, die verschiedene Zustände, und zwar sowohl Natur- als auch Körperzustände, charakterisieren, z.B.: *wild, roh, fest, frisch, leer, rein, schwach* u.a.

7. Zur LSU „**Vergleichende Charakteristika**" zählten wir Adjektive, die den Vergleich des bezeichneten Gegenstandes mit einem anderen vorsieht, z.B.: *ideal, mindest…, gleich, höchst…, relativ* u.a.

8. Die LSU „**Menschliche Eigenschaften**" enthält emotionale und intellektuelle Charakteristika, z.B.: *gemein* u.a.

9. Zur LSU „**Äußeres**" zählten wir Adjektive, die den Körperbau des Menschen, die Funktion ihrer Organe charakterisieren, z.B.: *blind, stumm* u.a.

10. Zur LSU „**Sachliche Lexik**" gehören Adjektive, die vom sachlichen und wissenschaftlichen Stil zeugen, z.B.: *adsorptiv, azeotrop, kriminal* u.a.

11. Unter den Einheiten der LSU „**Zugehörigkeit**" – Adjektive zur Bezeichnung der (Nicht-)Zugehörigkeit usw., z.B.: *sozial, national, eigen, privat, frei, fremd* u.a.

12. Die LSU „**Physische Eigenschaften**" schließt Adjektive zur Bezeichnung der Form, Farbe, des Geschmacks, Klangs, Geruchs, der Temperatur, des Gewichts, der Geschwindigkeit, z.B.: *weiß, rot, sauer, spitz, schwarz, trocken, steil* u.a. ein.

13. Die LSU „**Material**" enthält Adjektive, die den Gegenstand bezüglich des Materials, aus dem es hergestellt ist, charakterisieren, z.B.: *silbern* u.a.

14. Zur LSU „**Ort**" gehören Adjektive, die den Gegenstand aus der Sicht seiner Lokalisierung oder seiner räumlichen Unterbringung charakterisieren, z.B.: *ober…, mittler…, hinter…, unter…, inner…* u.a.

Die semantischen Charakteristika von Verben erfassen Verben der Aktion, des Zustands und Prozesses sowie Modalverben, die als Bestimmungswörter in Substantivkomposita nicht festgestellt wurden. Behandeln wir die erhaltenen Unterklassen präziser:

1. Die LSU „**Bewegung**" enthält Verben zur Bezeichnung der aktiven und passiven Bewegung sowie der Veranlassung (des Antriebs) zur Bewegung, z.B.: *fahren, drehen, laufen, sprühen, fallen, fließen, reiten, schleppen, tanzen* u.a.

2. Zur LSU „**Geräusche und Kommunikation**" gehören Verben zur Bezeichnung des Schallens sowie der menschlichen und tierischen Kommunikation, z.B.: *sprechen, streiten, schmähen, reden, werben, nennen, schimpfen, schweigen* u.a.

3. Zur LSU „**Allgemeine Handlung**" gehören Verben zur Bezeichnung einer allgemeinen Handlung, z.B.: *leiten, spielen, turnen* u.a.

4. Unter den Einheiten der LSU „**Verhalten**" – Verben zur Bezeichnung von Handlungen und Aktionen des Menschen, z.B.: *prüfen, tarnen, fördern, strafen, wetten, fasten, stören* u.a.

5. Die LSU „**Zielerreichung**" enthält z.B.: *nutzen* u.a.

6. Die LSU „**Bearbeitung des Gegenstands**" umfasst Verben, die den physischen und intellektuellen Einfluss auf das Objekt bezeichnen, z.B.: *waschen, bauen, kochen, baden, backen, braten, heizen, räuchern, messen, spülen* u.a.

7. Zur LSU „**Physiologische Handlung**" gehören Verben zur Bezeichnung der Funktionen verschiedener Organe von Menschen und Tieren, z.B.: *schlafen, essen, speisen, trinken* u.a.

8. Zur LSU „**Übergabe und Erhalten**" zählen wir Verben der auf das Objekt gerichteten Handlung mit der Absicht des Besitzerwechsels, z.B.: *sammeln, wechseln, sparen, kaufen, opfern* u.a.

9. Die LSU „**Physischer Einfluss auf ein Objekt**" enthält Verben der Zerstörung, des Schaffens, Wechsels, die auf das Objekt gerichtet sind, z.B.: *trennen, sprengen, schlagen, brechen, klappen, kratzen, pressen, schneiden, wickeln* u.a.

10. Zur LSU „**Zustand**" gehören Verben des Zustands, des temporären Zustands und des Zustandswechsels sowie des beabsichtigten Zustands (Kausation), z.B.: *schalten, sieden, platzen, schlachten, tiefkühlen* u.a.

11. Zum Satz des LSU „**Moralischer Einfluss**" gehören Verben des psychologischen Einflusses, z.B.: *betteln* u.a.

12. Die LSU „**Sensorische Wahrnehmung**" enthält Verben, die die Wahrnehmung bezeichnen, z.B.: *schauen, stinken, blicken, fernsehen, horchen, lauschen, sehen* u.a.

13. Zur LSU „**Gefühle**" gehören Verben zur Bezeichnung psychischer Reaktionen des Menschen, z.B.: *lachen, trauern* u.a.

14. Die LSU „**Geistige Arbeit**" schließt Verben zur Bezeichnung der intellektuellen Arbeit des Menschen ein, z.B.: *denken, lehren, rektifizieren, lernen* u.a.

15. Zur LSU „**Dasein**" zählten wir Verben zur Bezeichnung des organischen und anorganischen Daseins, z.B.: *sterben, brennen, leben* u.a.

16. Zur LSU „**Lage**" gehören Verben zur Bezeichnung der Lage von Objekten, z.B.: *wohnen, ruhen, warten, liegen* u.a.

17. Die LSU „**Besitz und Zugehörigkeit**" enthält Verben, die auf das Vorhandensein des Objekts zu jemandes Verfügung hinweisen, z.B.: *fehlen, halten* u.a.

18. Unter den Einheiten der LSU „**Prozess**" – Verben zur Bezeichnung von allgemeinen und konkreten Prozessen, z.B.: *tauen* u.a.

19. Die LSU „**Phasenverben**" enthält Verben zur Bezeichnung von Perioden, Stadien, Etappen des Prozesses, z.B.: *dauern* u.a.

Auf solche Weise wurden die substantivischen Grundwörter in den drei untersuchten Modellen N + N (z.B. *Eintrittstemperatur, Augenblick, Bürgermeister*), A + N (z.B. *Sozialdemokrat, Feinkorn, Idealverhalten*), V + N (z.B. *Nutzpflanze, Schreibtisch, Wohngeld*) nach 34 lexikalisch-semantischen Unterklassen und die Bestimmungswörter in den obengenannten Modellen nach 34 Unterklassen von Substantiven, 14 Unterklassen von Adjektiven und 19 Unterklassen von Verben eingeteilt.

4. Haupteigenschaften der Kombinierbarkeit von Konstituenten eines Kompositums

Die Kombinierbarkeit von Konstituenten, aus denen ein Substantivkompositum besteht, kann

man nach einigen Eigenschaften charakterisieren. Zu den wichtigsten von ihnen gehören unserer Meinung nach solche wie die Breite und die Selektivität der Kombinierbarkeit, ihre Intensität und die Anzahl von signifikanten Verbindungen.

1. Kombinierbarkeitsbreite/-selektivität (KB/KS). In unserer Untersuchung werden, wie oben angeführt, 34 Unterklassen von Substantiven genannt, die als Grund- und Bestimmungswörter vorkommen. D.h., die Gesamtzahl von potenziellen Verbindungen, die diese Unterklassen bilden können, beträgt 34 × 34 = 1156. Aber, wie die Tabelle A (siehe Anhang) zeigt, sind nicht alle Zellen der Tabelle mit tatsächlich festgestellten Verbindungen gefüllt (in unserem Fall 943 Verbindungen). Die Eigenschaft der Kombinierbarkeit wird als „Kombinierbarkeitsbreite" (KB) bezeichnet. Für die Tabelle A gleicht dieser Wert KB = 943 ÷ 1156 = 0,8. Der gegensätzliche Begriff, der den „Schwung" der Kombinierbarkeit charakterisiert, ist ihre Selektivität. Je mehr Verbindungen für dieses Modell konstatiert sind, desto niedriger ist der Selektivitätsgrad und umgekehrt. Für die Tabelle A kann man die Selektivität als die relative Anzahl der unausgefüllten Zellen messen – KS = 213 ÷ 1156 = 0,2.

Wie die Berechnungen zeigen, sind die größte Breite und die kleinste Selektivität für das Modell N + N charakteristisch (KB = 0,8, KS = 0,2) und die kleinste Breite sowie den höchsten Grad der Selektivität hat das Modell V + N (KB = 0,4, KS = 0,6). Das Modell A + N wird mit dem „mittleren" Grad der Breite sowie der Selektivität charakterisiert (je 0,5).

2. Die Intensität der Kombinierbarkeit erhält man mit Hilfe des Kontingenzkoeffizienten, der seinerseits mit Hilfe des Chiquadrat-Tests berechnet wird. Je größer die Werte des Kontingenzkoeffizienten, desto stärker ist die Verbindung zwischen den Komponenten eines Kompositums. So ist die Intensität die Stärke der Verbindung zwischen den Komponenten.

3. Die Zahl von signifikanten Verbindungen der Unterklasse: Nicht alle Verbindungen zwischen den Komponenten, die in den Tabellen eingetragen sind, übertreffen nach ihrem quantitativen Wert eine bestimmte theoretisch zu erwartende Größe. Z.B. ist das semantische Modell LSU der Verben „Bearbeitung des Gegenstandes" + LSU der Substantive „Gegenstände und Instrumente" (z.B. *Backofen, Klebekachel, Okuliermesser, Spülmaschine*) in der Belletristik 29 Mal festgestellt (siehe Anhang C1). Spezielle Berechnungen zeigen, dass sich diese empirische Größe nicht signifikant von der theoretischen (berechneten) Größe (21,2) unterscheidet. Und umgekehrt ist in einigen Fällen der Unterschied zwischen der empirischen und der theoretischen Größe signifikant, z.B., das Modell LSU der Verben „Lage" + LSU der Substantive „Gebäude und Bauten" (z.B. *Stehquartier, Wohnzimmer, Wartesaal*), in dem tatsächlich 11 Substantive enthalten und theoretisch nur 2 zu erwarten sind. Das kann mit Hilfe des Chiquadrat-Tests für eine Zelle festgestellt werden, in dem diese Begriffe „theoretisch zu erwartende" und „empirische" Werte (Größen) gebraucht werden. Die Verbindungen, für die χ^2 den kritischen Wert von 3,84 überschreitet, werden auf der $\alpha = 0.05$-Ebene als signifikant betrachtet. Die Zahl von signifikanten Verbindungen in jeder Unterklasse kann nicht übereinstimmen.

Die letzten zwei Eigenschaften (Intensität und die Zahl von signifikanten Verbindungen) werden wir im Laufe des Beitrags noch ausführlicher behandeln.

5. Verbindungscharakteristika der Konstituenten in den Modellen N + N, A + N, V + N

Bei der Untersuchung der Kombinierbarkeit von semantischen Substantivunterklassen in einem Kompositum diente uns das Hauptwort als Grundlage, denn es „ist die semantische Basis der ganzen Konstruktion" (Naumann 2000: 43); deswegen erforschten wir es aus der Sicht der quantitativen und qualitativen Besonderheiten der Kombinierbarkeit mit dem Bestimmungswort. Die Zahlenangaben der Verteilung von semantischen Modellen wurden mit dem Chiquadrat-Test und dem Kontingenzkoeffizienten analysiert, die nicht nur das Vorhandensein

der signifikanten Verbindung zwischen den Komponenten eines Kompositums, sondern auch ihre Stärke feststellen lassen. Das Gesamtergebnis ist in der Tabelle D (siehe Anhang) wiedergegeben (in der Tabelle sind die Werte der stärksten Verbindung für das Modell hervorgehoben).

Als Beispiel führen wir eine alternative (Vierfelder)Tabelle für die Berechnung des χ^2 für das Modell LSU der Substantive „Person" + LSU der Substantive „Person" (z.B. *Zwillingsbruder, Partisanenjäger, Sklavenmädchen*) (siehe Tabelle 1).

Tabelle 1
Die Verteilung der LSU „Person" als Grund- und Bestimmungswort

Grundwort / Bestimmungswort	LSU „Person"	Andere LSU	Insgesamt
LSU „Person"	76 (a)	(b) 528	604
Andere LSU	613 (c)	(d) 6189	6802
Insgesamt	689	6717	7406

Das Chiquadrat für die Tabelle ist gleich 8,44. Da die Freiheitsgrade *df* für diese Tabelle *df* = 1 betragen und die kritische Größe $\chi^2 = 3{,}84$, ist die erhaltene Summe höher als die kritische. Um den Grad der Abhängigkeit (der Verbindung) zwischen den Komponenten dieses Modells zu berechnen, verwenden wir die Formel für den Koeffizienten *Φ*.

Im oben angeführten Beispiel gleicht $\Phi = +0{,}03$. Das bedeutet, dass im semantischen Modell LSU „Person" + LSU „Person" zwischen den Komponenten eine signifikante Verbindung besteht.

Analog berechneten wir die Werte für andere semantische Modelle. Diejenigen von ihnen, die signifikante Verbindungen aufweisen ($\chi^2 > 3{,}84$), haben Kontingenzkoeffizienten von 0,27 (das Modell LSU „Pflanzen" + LSU „Pflanzen" (z.B. *Beerenobst, Blumenkohl, Gemüsekürbis*)) bis 0,02 (insgesamt 13 Modelle). Nach der Berechnung der Durchschnittswertsgröße der signifikanten Verbindungen haben wir festgestellt, dass sie 0,05 beträgt. D.h., wir können folgende Verbindungstypen zwischen den Komponenten der Substantivkomposita mit dem Modell N + N aussondern:

1) Modelle mit starker Verbindung – $K > 0{,}05$;
2) Modelle mit mittlerer Verbindung – $K = 0{,}05$;
3) Modelle mit schwacher Verbindung – $K < 0{,}05$.

Diese Daten sind in der Tabelle 2 angeführt (+ – starke Verbindung, Δ – mittlere, • – schwache).

Einige Modelle in Tabelle 2 verdienen eine besondere Aufmerksamkeit. Zwischen den Substantivunterklassen, die die Bestimmungs- und Hauptwörter repräsentieren, lässt sich eine semantische Kongruenz beobachten. Das äußert sich vor allem darin, dass in jenen Fällen, in denen signifikante Verbindungen festgestellt werden (in der Tabelle 2 sind sie mit dem Zeichen + gekennzeichnet), sich die Unterklasse des Hauptwortes in jedem Fall mit derselben Unterklasse des Bestimmungswortes (LSU „Person" + LSU „Person" (z.B. *Sultanstochter, Frauenmörder, Königingemahl*), LSU „Somatismen" + LSU „Somatismen" (z.B. *Augenlid, Fußgelenk, Kopfhaut*), LSU „Attribute des Menschen" + LSU „Attribute des Menschen" (z.B. *Rockkragen, Stiefelhose*) usw.) verbindet. In den genannten Fällen kann man über die **semantische Identität** von Konstituenten eines Kompositums sprechen. In anderen Fällen beobachten wir eine **semantische Kongruenz** zwischen den Komponenten von Zusammensetzungen. So verbindet sich die Unterklasse „Person" mit der Unterklasse „Sammelbezeichnungen von Menschen, Organisationen" (z.B. *Arbeiterunion, Fabrikantenfamilie, Nachbarhorde*), die Un-

terklasse „Somatismen“ mit der Unterklasse „Physiologische Sphäre“ (z.B. *Herzkrankheit, Leibschmerz, Magenblutung*).

Tabelle 2

Signifikante Verbindungen zwischen den Komponenten der semantischen Modelle in den Substantivkomposita nach dem Modell N + N

2. / 1.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	•		•	•				•										•	Δ											Δ		
2		+	+			•				+																							
3			+						Δ		•									+				•									
4				+				•	Δ		•																						
5					+	+	•			+																							
6		•		+		+			+		Δ		•																	•			
7	•						+					•																					
8	•						•	+																									
9	•	•	•					•	+																								
10		•			+				+	+																							
11											•																				•			
12																					•				•					•			
13													•		•	•						•				•	•					Δ	
14											•			•					•		•				Δ	•						•	
15											•		•																				
16							•								Δ											•							•
17																										•							
18																		•	•		Δ						•						
19																			Δ							Δ						•	
20			•											•						+					•								
21																			•		•	•				•				•		•	•	
22																					•												
23						Δ																	+										•
24								•																									
25				•										•		•							+		+								
26											•					•		•			Δ					+							•	
27							•																		+		•							
28	•						•					•							•				•										
29	+																												•			Δ		
30																					•								•	+		•		
31																						Δ				Δ					+		+	
32	+																												Δ			+		
33																		•	+															
34	•																																	+

* 1. – N LSU des Bestimmungswortes, 2. – N LSU des Grundwortes.

Eine wichtige Beobachtung, die man aufgrund der Tabelle A (siehe Anhang) machen kann, ist die Tatsache, dass nicht alle Zellen der Tabelle gefüllt sind. Der Selektivitätsgrad, wie oben gezeigt, ist das Komplement zu der relativen Zahl der gefüllten Positionen. Von den

für die Tabelle A möglichen 1 156 Positionen sind nur 943 gefüllt. D.h., der Selektivitätsgrad für die Verbindungen im Modell N + N gleicht – KS = 0,2.

Die semantische Kombinierbarkeit der Komponenten im Modell N + N für verschiedene Stile enthält einige Besonderheiten. So ist in der Belletristik die stärkste Verbindung für die Modelle LSU „Pflanzen“ + LSU „Pflanzen“ (K = 0,27) (z.B. *Kokospalme, Pfirsichbaum, Weizenkorn*), LSU „Tiere“ + LSU „Somatismen“ (K = 0,21) (z.B. *Dorschkopf, Kuhauge, Rinderherz*) sowie LSU „Zeit, Alter“ + LSU „Zeit, Alter“ (K = 0,20) (z.B. *Herbsttag, Jahreszeit, Sommerabend*) charakteristisch. Im allgemeinen wies die Belletristik 98 semantische Modelle mit einer signifikanten Verbindung zwischen den Komponenten auf. Im publizistischen Stil wurde eine kleinere maximale Verbindungsstärke zwischen den Komponenten der semantischen Modelle gefunden, nämlich 0,15. Gerade solche Verbindungen zeigen die Modelle LSU „Gebäude und Bauten“ + LSU „Gebäude und Bauten“ (z.B. *Bunkereingang, Hotelterrasse, Scheunentor*), LSU „Zeit, Alter“ + LSU „Dasein“ (z.B. *Wintersanfang, Zeitgeschehen*) und LSU „Zeit, Alter“ + LSU „Zeit, Alter“ (z.B. *Jahresfrist, Morgenstunde*). In diesem Stil haben wir 21 Modelle mit einer signifikanten Verbindung zwischen Komponenten gefunden. Für den wissenschaftlichen Stil sind die folgenden Verbindungen am charakteristischsten: LSU „Pflanzen“ + LSU „Pflanzen“ (K = 0,26) (z.B. *Efeublatt, Kartoffelblatt, Safranreis*), LSU „Essen und Getränke“ + LSU „Pflanzen“ (K = 0,26) (z.B. *Butterpilz, Nahrungspflanze*) und LSU „Person“ + LSU „Sammelbezeichnungen von Menschen, Organisationen“ (K = 0,21) (z.B. *Christenunion, Kinderbande, Schützenverein*). Die Gesamtzahl der Modelle mit signifikanter Verbindung im wissenschaftlichen Stil ist 34.

14 LSU der Adjektive und 34 LSU der Substantive können theoretisch 476 semantische Modelle bilden. In unserer Untersuchung wurden de facto etwas mehr als die Hälfte davon gefunden, nur 240 (siehe Tabelle B im Anhang). Der Selektivitätsgrad der Kombinierbarkeit in diesem Modell beträgt also KS = 0,5.

Zu den häufigsten Modellen – je 17 Einheiten – gehören die LSU der Adjektive „Ort“ + LSU der Substantive „Somatismen“ (z.B. *Vorderbein, Oberhand, Unterlippe*) und LSU der Adjektive „Ort“ + LSU der Substantive „Raum und Ort“ (z.B. *Lokalebene, Hinterland, Mittelmeer*). Bei 99 Verbindungen ist nur je 1 Kompositum belegt.

Der Kontingenzkoeffizient K für die erhaltenen Modelle schwankt von 0,27 bis 0,08. Indem wir alle K-Werte addierten und diese Zahl durch ihre Anzahl dividierten, erhielten wir den Durchschnittswert von $K - K$ = 0,16, der den mittleren Grad der Verbindung in diesem Modell charakterisiert. Auf solche Weise haben wir die folgende Skala des Verbindungsgrades zwischen den Komponenten im semantischen Modell für die Konstruktion A + N erhalten:

1) Modelle mit starker Verbindung – K > 0,16;
2) Modelle mit mittlerer Verbindung – K = 0,16;
3) Modelle mit schwacher Verbindung – K < 0,16.

Die Modelle mit signifikanter Verbindung sind in der Tabelle 3 dargestellt.

Wie auch im Modell N + N beobachten wir im Modell A + N die semantische Kongruenz zwischen den Komponenten der Verbindungen. Zu solchen Modellen können wir die LSU der Adjektive „Zustand“ + die LSU der Substantive „Stoffe und Materialien“ (z.B. *Rohgas, Feststoff, Reinkautschuk*), die LSU der Adjektive „Vergleichende Charakteristika“ + die LSU der Substantive „Kennwerte und Eigenschaften der Gegenstände“ (z.B. *Relativgeschwindigkeit, Gleichgewicht*), die LSU der Adjektive „Ort“ + die LSU der Substantive „Raum und Ort“ (z.B. *Innenstadt, Oberland, Unterwelt*) zählen. Dabei ist es klar, dass es keine volle Identität für die Wörter geben kann, deren Komponenten Adjektive und Substantive sind.

Tabelle 3
Signifikante Verbindungen zwischen den Komponenten der semantischen Modelle in zusammengesetzten Substantiven mit der Konstruktion A + N

Grundwort / Bestimmungswort	1	2	3	5	6	7	8	9	10	13	19	22	26	29	32	33
1								•							•	
2																•
3	•											Δ				
6					Δ											
7											•		Δ			
10										Δ						
11											•			Δ		
12		•		Δ	•				•							
14			Δ			Δ	•									

\+ - starke Verbindung, Δ - mittlere, • - schwache

Die Kombinierbarkeit der Komponenten im Modell A + N in verschiedenen Stilen zeigt bestimmte Besonderheiten. In der Belletristik gibt es insgesamt 11 semantische Modelle mit signifikanter Verbindung zwischen den Komponenten. Der höchste Verbindungsgrad charakterisiert das Modell der LSU der Adjektive „Physische Eigenschaften“ + der LSU der Substantive „Pflanzen“ ($K = 0{,}26$) (z.B. *Weißdorn, Rotklee*) und der LSU der Adjektive „Ort“ + der LSU der Substantive „Somatismen“ ($K = 0{,}21$) (z.B. *Hinterkopf, Oberkiefer*). In der Publizistik haben wir nur 3 signifikante Verbindungen zwischen den Komponenten von semantischen Modellen festgestellt: die stärksten – im Modell LSU der Adjektive „Zugehörigkeit“ + LSU der Substantive „Staat, seine Attribute“ ($K = 0{,}41$) (z.B. *Privatwirtschaft, Imperial-Kaisertum*) und LSU der Adjektive „Ort“ + LSU der Substantive „Raum und Ort“ ($K = 0{,}38$) (z.B. *Untergrenze, Oberland*). 3 signifikante Verbindungen ergibt auch der wissenschaftliche Stil. Hier ist der höchste Verbindungsgrad im Modell LSU der Adjektive „Vergleichende Charakteristika“ + LSU der Substantive „Kennwerte und Eigenschaften der Gegenstände“ ($K = 0{,}38$) (z.B. *Gleichgewicht, Idealform*).

19 semantische Unterklassen von Verben und 33 semantische Unterklassen von Substantiven (LSU „Eigennamen“ fehlt) ergeben theoretisch die Möglichkeit der Bildung von 627 semantischen Modellen solcher Einheiten. Wir haben nur 275 gefunden. So ist der Selektivitätsgrad der Kombinierbarkeit für das Modell V + N gleich 0,6. Wir können annehmen, dass für die semantischen Modelle dieser Verbindung Homogenität kennzeichnend ist.

Die gefundenen Modelle sind in Tabelle C (siehe Anhang) zu sehen. Der größte Umfang kennzeichnet Modelle LSU der Verben „Bearbeitung des Gegenstandes“ + LSU der Substantive „Gegenstände und Instrumente“ (z.B. *Waschbalge, Kochgeschirr, Klebekachel, Ladestock*) sowie LSU der Verben „Bewegung“ + LSU der Substantive „Gegenstände und Instrumente“ (27) (z.B. *Rennauto, Paddelboot, Laufmasche, Fallschirm, Reitsitz, Fahrkarte*). Zugleich sollte man bemerken, dass 115 Modelle nur je einmal gebraucht wurden.

Die statistischen Berechnungen zeigen *K*-Werte von 0,08 bis 0,23. Indem wir alle erhaltenen Werte addiert und deren Summe durch ihre Anzahl dividiert haben, haben wir den Durchschnittswert *K* für diese semantischen Modelle festgestellt, nämlich 0,13.

Auf diese Weise können wir die signifikanten Verbindungen zwischen den Komposita-Konstituenten mit dem Modell V + N in folgende Typen unterteilen:

1) Modelle mit starker Verbindung – $K > 0{,}13$;

2) Modelle mit mittlerer Verbindung – K = 0,13;
3) Modelle mit schwacher Verbindung – K < 0,13.

Die Modelle mit einer signifikanten Verbindung und ihre Werte sind in der Tabelle 4 angegeben.

Tabelle 4
Signifikante Verbindungen zwischen den Komponenten der semantischen Modelle in zusammengesetzten Substantiven mit der Konstruktion V + N

Grundwort / Bestim-mungswort	1	5	6	8	13	18	19	25	26	32
1	•									
2							+	•		
4										•
6		•	Δ							
7				•		+				
8										+
9					•					
12					+					
14									•	
16				+						

+ - starke Verbindung, Δ - mittlere, • - schwache

Wie bei den vorigen Konstruktionen handelt es sich auch hier um die Kongruenz der Komponenten einiger Modelle, im Besonderen bei LSU der Verben „Geräusche und Kommunikation“ + LSU der Substantive „Sprache und Rede“ (z.B. *Schmähgeschrei, Rufname, Lobrede, Streitschrift*) und LSU der Verben „Physischer Einfluss auf das Objekt“ + LSU der Substantive „Tätigkeit, Aktion“ (z.B. *Spleißarbeit, Wühltätigkeit, Spanndienst*).

Wenn wir die signifikanten Verbindungen zwischen den Komponenten des Modells V + N gesondert für jeden Stil vergleichen, so sehen wir einen beträchtlichen Unterschied in ihrer Verteilung. Die Belletristik, in der insgesamt 13 semantische Modelle eine signifikante Verbindung zwischen den Komponenten zeigen, wird die höchste Signifikanz in den Modellen LSU der Verben „Lage“ + LSU der Substantive „Gebäude und Bauten“ (z.B. *Wohnbarack, Stehquartier, Liegestütze, Wartehalle, Ruheraum*) (K = 0,27), LSU der Verben „Schallen und Kommunikation“ + LSU der Substantive „Sprache, Rede“ (z.B. *Lobrede, Schmähgeschrei, Werbeplakat, Schimpfwort*) (K = 0,22) und LSU der Verben „Übergabe und Erhalten“ + LSU der Substantive „Sammelbezeichnungen von Menschen, Organisationen“ (z.B. *Mietpartei, Kaufleute, Sparverein*) (K = 0,22) erreicht. In der Publizistik ist es nicht gelungen, eine Analyse durchzuführen, denn keines von den festgestellten Modellen wird mehr als dreimal gebraucht. Im wissenschaftlichen Stil repräsentieren zwei Modelle mit einer signifikanten Verbindung folgende Verbindungswerte: LSU der Verben „Bearbeitung des Gegenstandes“ + LSU der Substantive „Abstrakte Begriffe“ (z.B. *Heizzweck, Kühlmittel, Meßmittel*) – K = 0,30, LSU der Verben „Bearbeitung des Gegenstandes“ + LSU der Substantive „Stoffe und Materialien“ (z.B. *Destillierblase, Waschflüssigkeit, Baustein, Stickstoff*) – K = 0,27.

6. Vergleichscharakteristiken von semantischen Unterklassen der Grundwörter in den Modellen N + N, A + N, V + N

Die untersuchten Strukturmodelle N + N, A + N, V + N, die eine gemeinsame Konstituente – Substantiv als Grundwort – haben, zeigten in ihrem Funktionieren einige Unterschiede. Um zu bestimmen, wie groß die Ähnlichkeit ist, die den Gebrauch von verschiedenen semantischen Unterklassen dieser Komponente in Modellen mit verschiedenen Bestimmungswörtern (Substantiven, Adjektiven, Verben) charakterisiert, wurden die Angaben ihrer quantitativen Verteilung und deren Rangverteilung in eine allgemeine Tabelle eingetragen (siehe Tabelle G im Anhang).

Die Resultate zeigen, dass die Verteilung der Substantive nach den LSU ziemlich ungleichmäßig ist. Nur in 9 Fällen wurde eine Ähnlichkeit im Gebrauch von einzelnen Unterklassen festgestellt, und zwar bei LSU „Attribute des Menschen“, LSU „Pflanzen“, LSU „Raum und Ort“, LSU „Essen und Getränke“, LSU „Bewegung“, LSU „Tätigkeit, Aktion“, LSU „Sprache“, LSU „Naturerscheinungen und Zustände“, LSU „Eigennamen“.

Die Fälle, die uns Unterschiede im Gebrauch zeigen, werden wir präziser für jedes Wortbildungsmodell im Einzelnen betrachten. Zur Veranschaulichung der Verteilung von LSU der substantivischen Grundwörter in verschiedenen Strukturmodellen bringen wir Graphiken, wobei die *x*-Achse die LSU zeigt und die *y*-Achse den Rang der LSU (siehe Abbildung 1, 2, 3).

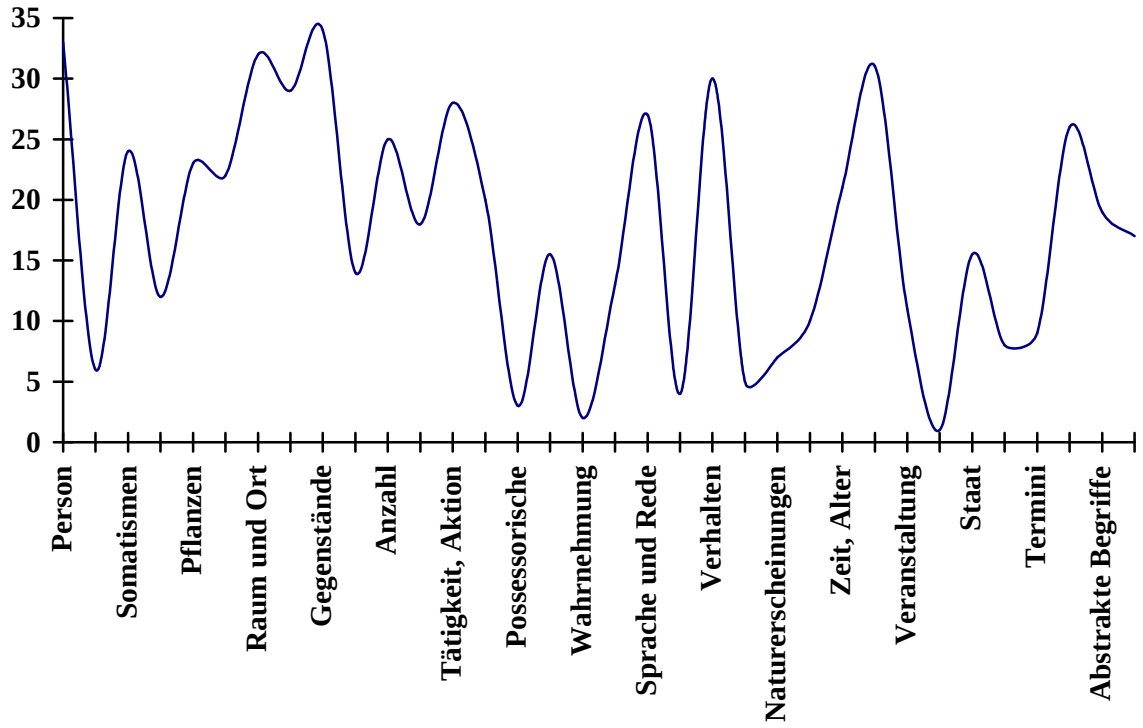


Abbildung 1. Gebrauch der LSU von substantivischen Grundwörtern im Wortbildungsmodell N + N

Im Vergleich zu anderen Konstruktionen werden im Wortbildungsmodell N + N als Grundwörter häufiger die Einheiten der LSU „Verhalten und Handlungen“ gebraucht und viel seltener die Einheiten der LSU „Tiere“ und „Eigenschaften des Menschen“.

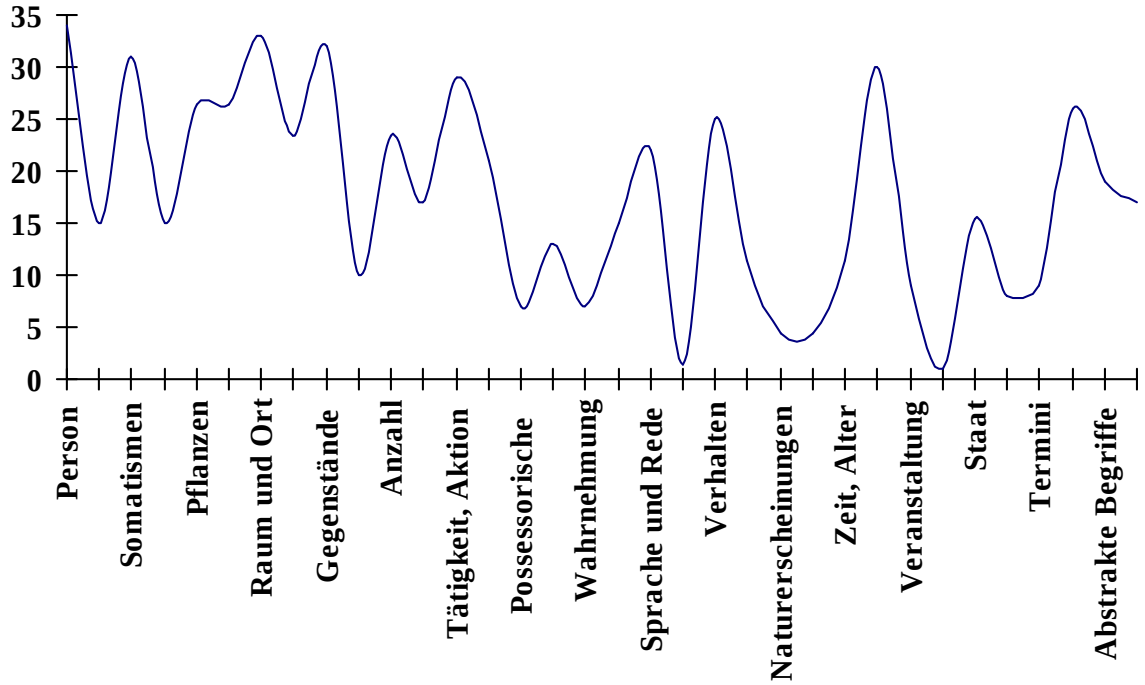


Abbildung 2. Gebrauch der LSU von substantivischen Grundwörtern im Wortbildungsmodell A + N

Die substantivischen Grundwörter, die sich in einem Kompositum mit einem Adjektiv verbinden, gehören häufiger als in anderen Wortbildungsmodellen zu den LSU „Somatismen“, „Possessorische Sphäre“, „Wahrnehmung“, „Staat, seine Attribute“; viel seltener kommen in dieser Rolle die Einheiten der LSU „Gebäude und Bauten“, „Physikalische Eigenschaften“, „Zeit, Alter“ vor.

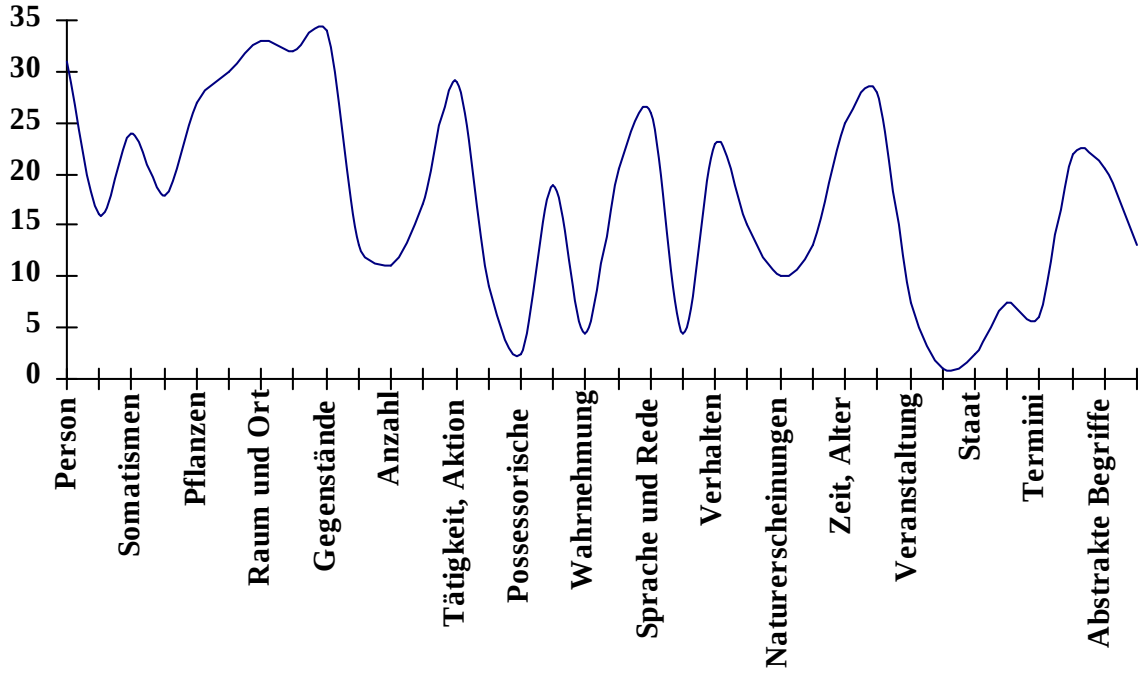


Abbildung 3. Gebrauch der LSU von substantivischen Grundwörtern im Wortbildungsmodell V + N

Die Grundwörter im Wortbildungsmodell V + N gehören häufiger zur LSU „Seelische Sphäre“. Etwas seltener werden in Verbindung mit dem verbalen Bestimmungswort Substantiv-Grundwörter der LSU „Anzahl, Maßeinheiten“, „Dasein“, „Staat, seine Attribute“ beobachtet.

Die entsprechenden Verteilungen der LSU von substantivischen Grundwörtern für verschiedene Wortbildungsmodelle offenbaren also sowohl Verschiedenheit als auch Ähnlichkeit. Um zu bestimmen, wie groß die Ähnlichkeit zwischen den Komponenten ist, haben wir die Korrelationsanalyse angewendet. Deren Ergebnisse gibt die Tabelle 5 an.

Tabelle 5
Ähnlichkeit bei den Wortbildungsmodellen N + N, A + N, V + N
im Gebrauch der LSU von Substantiv-Grundwörtern (Korrelationskoeffizient)

	N+N	A+N	V+N
N+N	–	0,881	0,848
A+N		–	0,650
V+N			–

Die erhaltenen Resultate bezeugen positive Korrelation in allen Paaren (bei $df = 32$ ist der minimale signifikante Koeffizient $P_{0,05} = 0,34$). Dabei ist die Ähnlichkeit im Gebrauch der semantischen Unterklassen von substantivischen Grundwörtern zwischen den Modellen mit Adjektiven und Verben viel geringer (0,650) als in den Modellen N + N und A + N (0,881), bzw. N + N und V + N (0,848).

7. Schlussfolgerungen

Ein Fazit aus der durchgeführten Untersuchung ziehend können wir feststellen, dass die Kombinierbarkeit der LSU, deren Einheiten als Bestimmungs- und als Grundwörter vorkommen, durch unterschiedliche Breite und Intensität gekennzeichnet wird. Die Verbindungen, in denen die Verbindungsstärke der Komponenten die Erwartungswerte in einem bestimmten Maße übertreffen, können wir als signifikant bezeichnen. Die Zahl der signifikanten Verbindungen für eine bestimmte LSU nennen wir semantische Produktivität der Unterklasse.

Die Bildung von semantischen Modellen in der Konstruktion N + N zeigt eine Reihe von Besonderheiten: So haben 17 LSU der Substantiv-Grundwörter die Tendenz zur Bildung von semantischen Modellen (mit der stärksten Verbindung) mit den Substantiv-Bestimmungswörtern derselben LSU. In diesem Fall können wir von einer Verschlossenheit dieser Unterklassen sprechen, denn das Bestimmungswort und das Grundwort werden unter den Einheiten derselben Unterklasse ausgewählt. Die anderen LSU zeigen dagegen Offenheit und verbinden sich am stärksten mit verschiedenen LSU. In diesem Fall bezeichnet die zweite Komponente meistens abstrakte Begriffe. Es darf also behauptet werden, dass die Substantive, die abstrakte Begriffe bezeichnen, offenere semantische Verbindungen innerhalb der Komposita eingehen und aktive Teilnehmer der Wortbildungsprozesse sind. Im Großen und Ganzen ist für die Kombinierbarkeit der lexikalischen Konstituenten in Komposita ein hoher Kongruenzgrad charakteristisch.

Die produktivsten Unterklassen hinsichtlich der Zahl von signifikanten Verbindungen sind die LSU der Grundwörter „Person“, „Kennwerte und Eigenschaften der Gegenstände“, „Sprache und Rede“, „Verhalten und Handlungen“ und „Tiere“.

Die Untersuchung der semantischen Modelle in der Konstruktion A + N bestätigte, dass für sie die Verbindung der Adjektive mit den Substantiven zur Bezeichnung von konkre-

ten Begriffen kennzeichnend ist. Die größte Zahl von signifikanten Verbindungen zwischen den Komponenten zeigen die LSU „Stoffe und Materialien“ und „Sprache und Rede“. Kennzeichnend für diese Konstruktionen sind Adjektive, die physische Eigenschaften und Orte bezeichnen.

Der höchste Grad der Verbindung zwischen den Komponenten kennzeichnet die semantischen Modelle LSU der Adjektive „Zugehörigkeit“ + LSU der Substantive „Staat, seine Attribute“, LSU der Adjektive „Physische Eigenschaften“ + LSU der Substantive „Pflanzen“, LSU der Adjektive „Zustand“ + LSU der Substantive „Stoffe und Materialien“, LSU der Adjektive „Vergleichende Charakteristika“ + LSU der Substantive „Kennwerte und Eigenschaften der Gegenstände.

Im Strukturmodell V + N hat die LSU „Gegenstände und Instrumente“ und die LSU „Raum und Ort“ die meisten signifikanten Verbindungen unter den Substantiv-Grundwörtern. Die stärkste Verbindung kennzeichnet die LSU „Gebäude und Bauten“, „Tätigkeit, Aktion“, „Sammelbezeichnungen von Menschen, Organisationen“. Die Verben zeigen eine große Zahl von signifikanten Verbindungen in den LSU „Bewegung“, „Bearbeitung des Gegenstands“ und „Physiologische Handlung“.

Unter den semantischen Modellen kennzeichnet die stärkste Verbindung die LSU der Verben „Lage + LSU der Substantive „Gebäude und Bauten“, LSU der Verben „Schallen und Kommunikation“ + LSU der Substantive „Sprache und Rede“, LSU der Verben „Übergabe und Erhalten“ + LSU der Substantive „Sammelbezeichnungen von Menschen, Organisationen“, LSU der Verben „Physiologische Handlung“ + LSU der Substantive „Seelische Sphäre“ und LSU der Verben „Sensorische Wahrnehmung“ + LSU der Substantive „Tätigkeit, Aktion“.

Die Vergleichsanalyse der LSU von Grundwörtern in unterschiedlicher Umgebung zeigte einen geringen Unterschied in der Verteilung der Substantive nach den LSU bei der Verbindung mit Substantiv-, Adjektiv- und Verb-Bestimmungswörtern. Die größte Ähnlichkeit beobachten wir im Gebrauch von semantischen Unterklassen der substantivischen Grundwörter zwischen den Modellen mit Substantiven und Adjektiven bzw. Substantiven und Verben.

Literatur

Altmann G. (1989). Hypotheses about Compounds. In: R. Hammerl (Hg.), *Glottometrika 10. 100-107.* Bochum: Brockmeyer.

Barz I. (2000). *Praxis- und Integrationsfelder der Wortbildungsforschung*. Heidelberg: Winter.

Bystrova L.V. (1978). Vyvčennja syntahmatyčnych zvjazkiv sliv za dopomohoju statystyčnych metodiv. *Movoznavstvo 4, 44-47.*

Chang Y. (2005). Anglizismen in der deutschen Fachsprache der Computertechnik. Eine korpuslinguistische Untersuchung zu Wortbildung und Bedeutungskonstitution fachsprachlicher Komposita. Frankfurt a. M.: Lang.

DW = Ortner L., Müller-Bollhagen E., Ortner H. u.a. (1991). *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Vierter Hauptteil: Substantivkomposita (Komposita und kompositionsähnliche Strukturen 1).* Berlin/New York: de Gruyter, 1991.

Donalies E. (2002). *Die Wortbildung des Deutschen: Ein Überblick*. Tübingen: Narr.

Dressler W. U. & Barbaresi L.M. (1991). Interradical interfixes: contact and contrast. In: V. Ivir & D. Kalogjera (ed.), *Languages in Contact and Contrast. Essays in Contact Linguistics: 133-145.* Berlin: Mouton de Gruyter.

Duden. *Die Grammatik der deutschen Gegenwartssprache, Bd. 4* (1995). Mannheim/Wien/ Zürich: Dudenverlag.

Erben J. (2000). *Einführung in die deutsche Wortbildungslehre. 4., aktualisierte und ergänzte Aufl.* Berlin: Erich Schmidt.

Feine A. (1997). Mit Spritfressern in die Klimakatastrophe? – Betrachtungen zu Mehrfachbenennungen in publizistischen Texten. In: Keßler Ch., Sommerfeldt K.-E. (Hgg.): *Sprachsystem – Text – Stil. Festschrift für Georg Michel und Günther Starke zum 70. Geburtstag; 61-74.* Frankfurt a.M.: Peter Lang.

Fleischer W. (1969). *Wortbildung der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut.

Fleischer W., Barz I. (1995). *Wortbildung der deutschen Gegenwartssprache*. Unter Mitarb. von Marianne Schröder., 2., durchges. und erg. Aufl. Tübingen: Niemeyer,

Gallmann P. (1999). Fugenmorpheme als Nicht-Kasus-Suffixe. In: Butt M., Fuhrhopp N. (Hg.) *Germanistische Linguistik 141-142, 177-190. Variation und Stabilität in der Wortstruktur.* Hildesheim u.a.: Georg Olms Verlag.

Henzen W. (1957). *Deutsche Wortbildung*. 2. verb. Aufl. Tübingen: Max Niemeyer Verlag.

Kluge F. (1913). *Abriss der deutschen Wortbildungslehre*. Halle: Verlag von Max Niemeyer, 1913.

Kühnhold I., Wellmann H. (1973). *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache – Band I „Das Verb“.* Düsseldorf: Schwann.

Kühnhold I., Putzer O., Wellmann H. (1978). *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache – Band III „Das Adjektiv“*. Düsseldorf: Schwann.

Langer S. (1998). Zur Morphologie und Semantik von Nominalkomposita. In: *Tagungsband der 4. Konferenz zur Vorbereitung natürlicher Sprache (KONVENS): 83-97.* Bonn.

Levkovskaja K.A. (1960). *Imennoje slovoobrayovanije v sovremennoj nemezkoj obščestvenno-političeskoj terminologii i primykajuščej k nej leksike*. Moskva: AN UdSSR.

Lindner Th. (1998). Zu Geschichte und Funktion von Fugenelementen in Nominalkomposita am Beispiel des Deutschen. *Moderne Sprachen 42,1-10.*

Motsch W. (1999). *Deutsche Wortbildung in Grundzügen*. Berlin/New York: de Gruyter.

Naumann B. (2000). *Einführung in die Wortbildungslehre des Deutschen*. 3., neubearb. Aufl. Tübingen: Niemeyer

Nübling D. (2004). Vom Name-n-forscher zum Name-ns-forscher: Unbefugte oder befugte ns-Fuge in Namen(s)-Komposita? In: Bok V. (Hgg.): *Studien zur deutschen Sprache und Literatur. Festschrift für Konrad Kunze zum 65. Geburtstag: 334-353.* Hamburg: Kovač.

Rings G. (2001). Wirtschaftskommunikation ohne Komposita und Derivate? Zur Vermittlung von Wortbildungsstrukturen in Theorie und Praxis des Wirtschaftsdeutschen. *German as a foreign language 1,1-28.* Online: http://www.gfl-journal.de/1-2001/rings.html.

Rothe U. (1988). Polylexy and Compounding. In: K.P. Schulz (Hg.), *Glottometrika 9, 121-134.* Bochum: Brockmeyer.

Sajenko A.N. (2002). Charakternyje osobennosti podjazykov spezial´nosti (na materiale nemezkogo jazyka). *Kul´tura narodov Pričernomorja 44, 176-180.*

Schröder M. (1979). Zu Beziehungen zwischen Wortbildung und Polysemie. *Deutsch als Fremdsprache 5, 286-291.*

Smirnova E.D. (1982). *Formalisirovannyje jazyki i problemy logičeskoj semantiki*. Moskva: MGU.

Steiner P. (2002). Polylexie und Kompositionsaktivität in Text und Lexik. In: Köhler R. (ed.). *Korpuslinguistische Untersuchungen in die quantitative und systemtheoretische Linguistik: 209-251.* http://ubt.opus.hbz-nrw.de/volltexte/2004/279/

Stepanova M. (1953). *Slovoobrasovanije sovremennogo nemezkogo jazyka*. Moskva: Izdatel´stvo literatury na inostrannych jazykach.
Vandermeeren S. (1999). Semantische Analyse deutscher Substantiv-Komposita: Drei Untersuchungsmethoden im Vergleich. *Leuvense bijdragen 88, 69-94*.

Quellenverzeichnis

Belletristik:

Böll H. (1995). *Fürsorgliche Belagerung.* München: Deutscher Taschenbuch Verlag.
Böll H. (2001). *Frauen vor Flußlandschaft*. München: Deutscher Taschenbuch Verlag
Grass G. (1980). *Kopfgeburten oder Sterben die Deutschen aus.* Berlin.
Grass G. (1999). *Der Butt*. München: Deutscher Taschenbuch Verlag
Heym S. (1974). *5 Tage im Juni*. Gütersloh: Reinhard Mohn OHG, 1974. – 384 S.
Heym S. (1987). *Schwarzenberg*. Frankfurt am Main: Fischer Taschenbuch Verlag.
Kant H. (1977). *Der Aufenthalt.* Berlin: Rütten und Loening.
Kant H. (1981). *Das Impressum*. Berlin: Rütten und Loening.
Lenz S. (1985). *Exerzierplatz.* Hamburg: Hoffmann und Campe.
Lenz S. (2001). *Arneß Nachlaß.* Munchen: Deutscher Taschenbuch Verlag.
Loest E. (1980). *Swallow, mein wackerer Mustang*. Hamburg: Hoffmann und Campe.
Loest E. (1984). *Völkerschlachtsdenkmal*. Hamburg: Hoffmann und Campe.
Walser M. (1987). *Brandung*. Frankfurt am Main: Suhrkamp Taschenbuch Verlag.
Walser M. (2002). *Seelenarbeit*. Frankfurt am Main: Suhrkamp Taschenbuch Verlag.
Wolf Ch. (1995). *Kindheitsmuster*. München: Deutscher Taschenbuch Verlag.
Wolf Ch. (2002). *Kassandra*. München: Deutscher Taschenbuch Verlag.

Wissenschaftliche Literatur:

Reinbothe H., Wasternack C. (1986). *Mensch und Pflanze. Kulturgeschichte und Wechselbeziehung.* 1. Aufl. Heidelberg/Wiesbaden: Quelle & Meyer Verlag.
Steinbuch P. A. (1993). *Fertigungswirtschaft.* 5., überarb. und erw. Auflage. Ludwigshafen (Rhein): Kiel.
Mikl-Horke G. (1992). *Soziologie: historischer Kontext und soziologische Theorie-Entwürfe.* 2., durchges. Aufl. München/Wien: Oldenbourg.
Ebeling N. (1999). Abluft und Abgas: Reinigung und Überwachung. In: Josef Kwiatkowski und Claus Bliefert (Hrsg.). *Praxis des technischen Umweltschutzes*. Weinheim/New York/Chichester/Brisbane/Singapore/Toronto: Wiley-VCH**.**
Kaempfert M. (1984). *Wort und Wortverwendung. Probleme der semantischen Deskription anhand von Beobachtungen an der deutschen Gegenwartssprache.* Göppingen: Kümmerle Verlag.
Rodingen H. (1977). *Pragmatik der juristischen Argumentation: was Gesetze anrichten u. was rechtens ist.* 1. Aufl. Freiburg (Br.)/München: Alber.
Sattler K. (1988). *Thermische Trennverfahren: Grundlagen, Auslegung, Apparate.* Weinheim/Basel (Schweiz)/Cambridge/New York: VCH.

Publizistik:

1. Die Welt, 1995
2. Die Zeit, 1988
3. Der Spiegel, 1994

Anhang

Tabelle A

Verteilung nach den semantischen Modellen in der Konstruktion N + N

Grundwort Bestim- mungswort	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	76	3	33	16	7	6	40	53	53	7	17	11	14	10	3	5	2
2	11	10	49	2	10	11	14	13	21	22	3	1	7	4	-	1	1
3	18	3	41	6	11	11	25	6	62	8	21	9	16	7	1	1	1
4	9	-	1	10	2	1	-	8	18	1	6	-	3	-	-	1	-
5	10	6	6	4	75	22	37	12	31	32	7	3	14	2	1	1	1
6	18	11	20	26	19	35	32	27	99	10	33	9	34	2	2	2	-
7	72	7	14	5	13	16	98	31	41	2	10	23	18	12	4	7	-
8	40	3	-	5	4	4	33	64	36	-	9	3	17	4	1	2	-
9	67	9	25	10	14	15	29	39	111	4	12	5	27	7	2	5	-
10	5	5	3	1	22	6	6	9	37	25	5	-	8	4	-	-	-
11	7	-	1	-	-	4	7	2	6	-	9	4	9	4	3	5	-
12	10	-	2	1	1	4	10	9	10	1	3	4	3	4	1	5	-
13	25	3	2	7	2	10	35	20	33	1	16	9	28	14	7	18	-
14	6	-	-	-	-	2	17	5	5	1	11	6	7	10	-	1	-
15	3	-	1	-	-	1	3	2	3	-	7	1	6	3	3	2	-
16	5	-	2	-	2	1	17	1	6	-	6	2	1	3	5	-	-
17	3	-	-	-	-	-	2	1	1	-	-	1	1	1	-	1	-
18	7	1	2	1	1	1	8	4	7	1	4	5	4	5	-	4	2
19	30	-	3	1	1	5	14	5	30	1	9	6	5	6	1	8	4
20	1	-	7	1	-	2	1	1	5	1	1	2	2	6	1	1	-
21	36	1	2	-	-	5	15	14	31	1	3	6	16	8	3	9	3
22	2	-	-	-	-	-	1	-	3	-	1	1	1	3	1	-	1
23	5	1	3	3	4	12	14	7	19	1	5	7	7	6	1	2	1
24	8	-	5	3	-	-	1	1	16	-	2	3	5	3	-	1	1
25	6	1	1	8	6	3	13	5	14	4	9	7	4	12	1	10	1
26	12	1	4	1	1	1	7	3	6	1	13	7	8	7	4	10	4
27	11	1	-	1	1	1	13	7	9	2	2	2	1	3	2	4	-
28	20	-	1	1	6	5	19	12	9	1	1	9	2	1	-	1	-
29	43	1	1	3	1	-	19	4	12	-	3	1	14	8	1	5	1
30	11	-	-	-	-	1	4	2	11	1	2	2	4	5	2	2	-
31	1	-	-	1	-	5	5	8	10	-	4	2	8	2	2	4	-
32	87	-	2	1	3	4	14	25	27	1	2	5	11	6	2	11	-
33	7	1	1	-	2	2	9	3	13	-	8	3	8	4	3	4	1
34	17	1	2	2	1	-	2	1	10	-	4	5	7	5	3	3	-
Σ	689	69	234	120	209	196	564	404	805	129	248	164	320	181	60	136	24

Fortsetzung der Tab. A

Grundwort Bestim-mungswort	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Σ
1	20	42	4	36	3	6	4	14	28	11	-	17	6	1	42	8	6	604
2	-	5	4	2	-	1	3	-	2	2	-	-	1	1	3	1	-	205
3	8	5	12	8	-	2	9	1	23	3	-	3	1	1	3	3	4	333
4	-	-	-	1	-	-	-	-	3	1	-	-	-	-	2	-	-	67
5	-	6	2	6	-	7	1	3	18	2	-	1	1	-	3	1	2	317
6	2	1	2	25	3	8	8	7	9	3	-	4	6	12	9	6	5	489
7	2	13	2	15	2	5	11	5	30	7	-	11	5	3	24	3	10	521
8	1	6	-	12	-	3	5	3	24	4	2	4	1	1	9	3	2	305
9	1	17	1	24	-	5	5	3	28	6	-	4	3	-	21	11	9	519
10	3	1	-	3	-	2	4	1	2	-	1	-	1	3	3	2	-	162
11	-	3	-	8	3	1	3	1	10	1	1	5	2	6	7	3	4	119
12	2	5	1	14	2	1	2	8	10	3	-	3	5	2	6	3	3	138
13	8	11	2	28	9	-	6	14	40	12	-	4	2	9	16	24	5	420
14	6	12	2	18	3	1	2	14	18	2	-	1	-	4	5	9	5	173
15	2	1	1	3	1	1	-	-	5	-	-	2	2	-	2	1	1	57
16	2	7	1	5	4	-	-	4	12	-	-	3	-	1	5	5	7	107
17	3	2	1	-	-	-	-	1	6	1	-	-	-	-	-	1	2	28
18	5	10	4	18	2	3	2	6	3	6	-	2	-	1	3	3	5	130
19	4	22	-	11	1	2	1	5	32	6	-	4	3	5	4	14	6	249
20	3	4	9	8	1	1	-	6	3	1	-	1	1	1	1	1	2	75
21	9	28	2	31	8	3	5	15	34	8	-	8	10	6	25	19	9	373
22	2	1	-	6	1	-	-	-	5	2	-	2	2	1	1	2	-	39
23	1	3	3	9	1	13	3	5	4	1	-	2	1	4	-	3	7	158
24	1	3	1	8	3	1	3	4	4	-	-	-	1	3	4	3	3	91
25	6	8	4	9	2	10	6	39	7	4	-	3	5	3	2	2	4	219
26	7	4	-	24	3	-	-	4	33	2	-	2	3	5	7	8	8	200
27	-	3	-	3	-	-	1	10	4	5	-	1	2	1	5	1	-	96
28	1	11	1	3	-	6	-	4	3	1	-	5	2	2	2	-	2	131
29	4	8	-	16	2	-	1	1	13	4	-	11	3	1	19	5	6	211
30	-	6	-	12	-	-	2	1	6	-	1	6	9	-	8	2	2	102
31	-	1	2	12	6	1	3	1	21	1	-	2	1	8	2	13	5	131
32	10	17	-	30	3	2	5	7	21	7	-	17	4	3	35	11	6	379
33	8	19	-	9	1	-	-	1	12	4	-	4	4	3	3	5	2	144
34	1	6	-	4	-	-	-	4	10	3	-	4	1	1	7	1	9	114
Σ	122	291	61	421	64	85	95	192	483	113	5	136	88	92	288	177	141	7406

* Die Grundwort- und Bestimmungswortnummer entspricht der Nummer der lexikalisch-semantischen Unterklasse von Substantiven

Tabelle B

Verteilung nach semantischen Modellen in der Konstruktion A + N

Grund-wort / Bestim-mungswort	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Σ
1	12	2	2	-	1	2	8	4	10	-	1	3	5	2	-	-	1	1	-	1	6	1	1	-	2	4	2	1	1	1	2	8	1	2	87
2	6	-	2	2	-	2	4	2	7	1	3	2	5	4	-	3	-	-	2	1	5	-	2	2	-	6	2	-	1	-	1	-	6	4	75
3	13	-	-	-	2	-	3	-	-	1	3	-	4	4	1	1	-	1	-	-	3	5	-	-	4	2	-	-	-	-	1	1	1	1	51
4	5	-	-	-	3	2	3	-	8	1	2	3	6	1	1	3	2	4	5	-	2	4	1	1	-	2	-	-	2	1	1	4	-	3	70
5	-	-	2	1	-	-	2	1	-	-	-	-	1	-	-	-	-	1	-	-	1	-	-	-	-	2	-	-	-	1	-	-	-	-	12
6	-	4	4	-	4	9	4	3	3	-	4	1	1	4	-	-	1	-	-	-	-	-	1	-	-	-	-	-	-	1	-	-	-	1	45
7	3	-	-	-	-	-	1	1	1	-	4	1	1	-	1	1	-	1	5	-	1	-	-	1	-	8	-	-	-	-	-	-	1	-	31
8	-	-	2	-	-	-	-	-	2	-	1	1	-	-	1	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	14
9	1	1	1	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1	-	-	7
10	-	-	-	-	-	2	-	-	-	-	2	-	5	1	-	1	-	-	2	-	-	-	-	1	-	1	-	-	-	-	-	2	1	-	18
11	15	-	2	1	1	-	3	4	3	2	-	1	1	3	1	1	-	2	8	-	7	1	-	-	1	7	4	1	13	2	2	3	4	5	98
12	6	5	9	4	15	9	7	2	2	5	1	-	5	1	3	-	-	1	2	-	1	-	-	-	1	2	-	-	-	-	-	4	-	-	85
13	-	1	2	4	1	-	-	-	4	-	-	-	-	-	-	-	1	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	15
14	12	-	17	1	1	2	17	9	5	-	3	2	-	1	-	1	-	1	1	1	1	-	-	1	3	2	1	1	1	-	1	7	2	1	95
Σ	73	13	43	13	28	28	52	26	45	10	26	14	34	21	8	12	8	13	25	3	27	11	7	7	11	36	9	3	18	6	8	30	17	18	703

Tabelle C

Verteilung nach semantischen Modellen in der Konstruktion V + N

Grund-wort / Bestim-mungswort	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	29	30	31	32	33	34	Σ
1	13	2	4	4	5	7	14	8	27	-	2	4	4	4	-	2	-	2	2	-	4	1	4	3	2	6	1	-	-	-	2	1	2	130
2	2	3	-	-	-	1	2	-	5	-	-	1	1	-	-	-	-	2	9	1	2	3	-	2	5	4	1	-	1	-	1	4	-	50
3	1	-	-	3	3	1	4	2	3	-	-	-	1	-	1	2	-	-	3	-	1	-	-	-	1	-	-	-	-	-	1	1	-	28
4	4	1	-	2	1	2	2	2	7	2	1	4	2	1	-	4	1	2	4	-	2	1	-	-	2	2	2	2	-	-	5	4	-	62
5	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	3
6	8	3	4	4	11	15	12	7	32	4	2	-	3	2	-	1	-	-	2	-	-	2	1	2	2	5	1	-	-	-	2	6	-	131
7	3	1	-	1	-	2	3	6	6	1	-	1	2	-	-	-	-	5	-	-	1	1	1	-	-	-	-	-	1	-	-	1	-	36
8	5	3	-	-	1	3	6	2	9	-	1	3	1	-	-	1	-	-	-	-	2	-	-	-	1	3	1	-	1	-	7	-	1	51
9	2	-	5	2	4	4	-	3	19	2	1	-	9	-	1	3	1	-	2	-	3	1	-	1	-	6	-	-	-	-	1	1	1	72
10	1	-	4	1	-	-	6	2	10	1	-	-	2	-	-	1	-	2	-	-	1	-	-	1	3	1	-	-	-	2	-	-	2	40
11	2	1	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6
12	1	-	3	-	1	-	4	3	4	-	1	-	6	-	-	-	-	-	-	-	2	1	-	-	1	1	-	-	-	-	-	1	-	29
13	1	-	3	1	1	-	1	-	1	-	-	1	-	-	-	-	1	3	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	14
14	1	-	-	-	1	2	2	3	10	-	1	-	4	-	-	2	-	2	3	-	4	2	-	1	4	6	-	-	2	1	-	-	4	55
15	1	-	1	-	2	2	-	-	9	2	-	-	1	-	-	-	-	1	3	-	-	1	-	1	3	-	-	-	-	-	-	1	-	28
16	-	1	2	-	-	-	5	11	4	-	1	-	-	-	-	-	-	-	-	1	1	-	-	-	4	1	-	-	-	-	1	-	1	33
17	-	-	-	-	-	-	2	1	-	-	-	-	1	-	-	2	-	2	1	-	1	-	-	1	-	-	-	-	1	2	1	-	1	16
18	-	-	-	-	-	-	-	-	1	-	-	2	-	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	5
19	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	1	-	-	-	1	-	-	2	-	1	-	-	-	-	-	-	-	-	6
Σ	45	15	26	18	31	39	64	51	148	12	10	17	37	7	2	19	3	21	30	3	24	13	9	12	29	36	6	2	6	5	22	21	12	795

Tabelle D
Signifikante Verbindungen zwischen den Komponenten von semantischen Modellen in Komposita mit der Konstruktion N + N (χ^2; K)

Grundwort / Bestim-mungswort	Person	Tiere	Somatismen	Attribute des Menschen	Pflanzen	Stoffe und Materialien	Raum und Ort	Gebäude und Bauten	Gegenstände und Instrumente	Essen und Getränke	Anzahl, Maßeinheiten	Bewegung
Person	8,44 0,03		11,46 0,04	4,39 0,02				14,12 0,04				
Tiere		35,55 0,07	296,28 0,20		3,24	5,95 0,03		0,32		99,49 0,12		
Somatismen			95,38 0,11	0,07	0,29	0,55			21,57 0,05	0,89	9,41 0,04	0,38
Attribute des Menschen	1,36			75,04 0,10				5,51 0,03	17,84 0,05		6,56 0,03	
Pflanzen		3,31			524,00 0,27	23,40 0,06	7,92 0,03			134,91 0,14		
Stoffe und Materialien		9,79 0,04	1,45	44,68 0,08	2,12	40,68 0,07			47,11 0,08	0,27	18,56 0,05	
Raum und Ort	13,64 0,04	1,04				0,37	97,93 0,12	0,27				12,58 0,04
Gebäude und Bauten	5,57 0,03						3,93 0,02	149,41 0,14	0,31			
Gegenstände und Instrumente	8,57 0,03	3,89 0,02	5,00 0,03	0,33		0,11		4,57 0,02	63,63 0,09			
Essen und Getränke		8,32 0,03			69,84 0,10	0,69			24,46 0,06	181,25 0,16		
Anzahl, Maßeinheiten											6,63 0,03	
Bewegung								0,31				
Tätigkeit, Aktion				0,01			0,36				0,29	
Dasein							1,28				4,95 0,03	1,28
Possessorische Sphäre											14,59 0,04	
Mentale Sphäre							10,70 0,04				1,71	
Wahrnehmung												
Seelische Sphäre												1,62
Sprache und Rede	2,29								0,37		0,06	0,04
Physiologische Sphäre			9,43 0,04									

Fortsetzung der Tab. D

Grundwort / Bestim-mungswort	Person	Tiere	Somatismen	Attribute des Menschen	Pflanzen	Stoffe und Materialien	Raum und Ort	Gebäude und Bauten	Gegenstände und Instrumente	Essen und Getränke	Anzahl, Maßeinheiten	Bewegung
Verhalten und Handlungen	0,05											
Eigenschaften des Menschen												
Naturerscheinungen und Zustände						15,37 0,05	0,41		0,25			3,72
Physikalische Eigenschaften			1,64						4,28 0,02			
Zeit, Alter				5,84 0,03							0,40	1,00
Kennwerte und Eigenschaften der Gegenstände											6,30 0,03	1,57
Veranstaltung, Spiel	0,53						4,94 0,03	0,63				
Eigennamen	5,61 0,03				1,5	0,69	9,13 0,04	3,54				13,34 0,04
Staat, seine Attribute	31,53 0,07						0,63					
Dokumente, Geld	0,27											
Termini						0,69		0,11				
Sammelbezeichnungen von Menschen, Organisationen	88,12 0,11							1,00				
Abstrakte Begriffe											2,21	
Wissenschaft, Kultur, Traditionen	4,31 0,02											2,52

* Die markierten Zellen zeigen die Modelle mit der stärksten signifikanten Verbindung.

Fortsetzung der Tab. D

Grundwort / Bestim-mungswort	Tätigkeit, Aktion	Dasein	Possessorische Sphäre	Mentale Sphäre	Wahrnehmung	Seelische Sphäre	Sprache und Rede	Physiologische Sphäre	Verhalten und Handlungen	Eigenschaften des Menschen	Naturerscheinun-gen und Zustände
Person						11,55 0,04	16,00 0,05		0,10		
Tiere											
Somatismen	0,20					1,28		32,96 0,07			
Attribute des Menschen											
Pflanzen	0,01										3,28
Stoffe und Materialien	8,68 0,03										1,08
Raum und Ort											
Gebäude und Bauten	1,23										
Gegenstände und Instrumente	1,04										
Essen und Getränke	0,15										
Anzahl, Maßeinheiten	3,07			3,75					0,24		
Bewegung				2,49					5,21 0,03		
Tätigkeit, Aktion	5,91 0,03	1,47	4,06 0,02	14,80 0,04		0,20			0,80	8,49 0,03	
Dasein		8,26 0,03				3,70	4,24 0,02		7,35 0,03		
Possessorische Sphäre	5,57 0,03										
Mentale Sphäre			20,14 0,05				1,96				
Wahrnehmung											
Seelische Sphäre		1,09				4,02 0,02	4,96 0,03		16,42 0,05		
Sprache und Rede				2,70			16,41 0,05				
Physiologische Sphäre		9,80 0,04						115,79 0,13	3,50		
Verhalten und Handlungen				0,72		1,48	13,29 0,04		5,04 0,03	7,51 0,03	
Eigenschaften des Menschen									6,87 0,03		

Fortsetzung der Tab. D

Grundwort / Bestim-mungswort	Tätigkeit, Aktion	Dasein	Possessorische Sphäre	Mentale Sphäre	Wahrnehmung	Seelische Sphäre	Sprache und Rede	Physiologische Sphäre	Verhalten und Handlungen	Eigenschaften des Menschen	Naturerscheinungen und Zustände
Naturerscheinungen und Zustände	0,01	1,27									71,87 0,10
Physikalische Eigenschaften	0,31								1,65		
Zeit, Alter		8,71 0,03		9,32 0,04		1,73					23,22 0,06
Kennwerte und Eigenschaften der Gegenstände		0,96		11,40 0,04		4,45 0,02			15,27 0,05		
Veranstaltung, Spiel											
Eigennamen							7,04 0,03				13,84 0,04
Staat, seine Attribute	2,81	1,65		0,34					1,45		
Dokumente, Geld		2,62					1,04		7,12 0,03		
Termini	1,03								3,00	21,48 0,05	
Sammelbezeichnungen von Menschen, Organisationen				2,51		2,50	0,32		3,70		
Abstrakte Begriffe	0,54					14,04 0,04	33,36 0,07		0,09		
Wissenschaft, Kultur, Traditionen	0,92	1,83					0,54				

Fortsetzung der Tab. D

Grundwort / Bestim-mungswort	Physikalische Eigenschaften	Zeit, Alter	Kennwerte und Eigenschaften der Gegenstände	Veranstaltung, Spiel	Eigennamen	Staat, seine Attribute	Dokumente, Geld	Termini	Sammelbezeichnungen von Menschen, Organisationen	Abstrakte Begriffe	Wissenschaft, Kultur, Traditionen
Person				0,39		3,51			16,59 0,05		
Tiere			0,08								
Somatismen	5,54 0,03										
Attribute des Menschen											
Pflanzen											
Stoffe und Materialien	0,51						0,01	8,50 0,03			
Raum und Ort	3,06		0,97			0,24			0,79		
Gebäude und Bauten	0,33										
Gegenstände und Instrumente									0,04		
Essen und Getränke			0,70								
Anzahl, Maßeinheiten			0,12			3,75		14,22 0,04	1,28		
Bewegung		5,71 0,03	6,56 0,03				7,09 0,03		0,08		
Tätigkeit, Aktion	0,07	0,96	4,37 0,02	5,24 0,03				2,94		21,07 0,05	
Dasein		21,20 0,05	0,53							6,00 0,03	0,95
Possessorische Sphäre			3,91 0,02								
Mentale Sphäre			10,23 0,04						0,18	2,42	8,08 0,03
Wahrnehmung											
Seelische Sphäre		2,14	16,91 0,05	8,40 0,03							2,72
Sprache und Rede				1,34				1,23		11,53 0,04	0,37
Physiologische Sphäre		8,76 0,03	4,32 0,02								
Verhalten und Handlungen	0,01	3,17	2,55	1,00		0,21	7,44 0,03	0,43	8,30 0,03	12,29 0,04	0,58

Fortsetzung der Tab. D

Grundwort / Bestimmungswort	Physikalische Eigenschaften	Zeit, Alter	Kennwerte und Eigenschaften der Gegenstände	Veranstaltung, Spiel	Eigennamen	Staat, seine Attribute	Dokumente, Geld	Termini	Sammelbezeichnungen von Menschen, Organisationen	Abstrakte Begriffe	Wissenschaft, Kultur, Traditionen
Eigenschaften des Menschen											
Naturerscheinungen und Zustände		0,22									5,70 0,03
Physikalische Eigenschaften											
Zeit, Alter	3,78	206,75 0,17	33,53 0,07				2,30				
Kennwerte und Eigenschaften der Gegenstände								2,65		2,28	4,93 0,03
Veranstaltung, Spiel		23,56 0,06		8,77 0,03					0,45		
Eigennamen						2,90					
Staat, seine Attribute						13,72 0,04			15,19 0,05		1,06
Dokumente, Geld			19,75 0,05			9,38 0,04	51,32 0,08		4,32 0,02		
Termini								25,70 0,06		32,42 0,07	2,66
Sammelbezeichnungen von Menschen, Organisationen			0,79	0,27		15,53 0,05			30,51 0,06	0,45	
Abstrakte Begriffe			0,96							0,74	
Wissenschaft, Kultur, Traditionen									1,57		22,49 0,06

Tabelle E

Signifikante Verbindungen zwischen den Komponenten von semantischen Modellen in Komposita mit der Konstruktion A + N (χ^2; K)

LSU des Grund-wortes / LSU des Bestim-mungswortes	Person	Tiere	Somatismen	Pflanzen	Stoffe und Materialien	Raum und Ort	Gebäude und Bauten	Gegenstände und Instrumente	Essen und Getränke	Tätigkeit, Aktion	Sprache und Rede	Verhalten und Handlungen	Eigenschaften des Menschen	Kennwerte und Eigenschaften der Gegenstände	Staat, seine Attribute	Sammelbezeichnungen von Menschen, Organisationen	Abstrakte Begriffe	Wissenschaft, Kultur, Traditionen
Größe, Abstand	1,24					0,47		4,30 0,08		0,18		2,51				5,90 0,09		
Anzahl								1,20		0,61		1,82		1,43			11,09 0,13	
Zeit	13,28 0,14												24,74 0,19					
Bewertung								3,28		2,36	2,92							
Zustand					32,25 0,21													
Vergleichende Charakteristika											14,95 0,15			28,56 0,20				
Sachliche Lexik										21,12 0,17								
Zugehörigkeit	2,96										7,05 0,10	3,36		0,96	52,30 0,27			2,95
Physische Eigenschaften		8,67 0,11	3,37	47,21 0,26	11,03 0,13	0,10			13,72 0,14	0,23								
Ort	0,60		26,54 0,19			17,67 0,16	10,29 0,12									2,59		

* Die markierten Zellen zeigen die Modelle mit der stärksten signifikanten Verbindung.

Tabelle F

Signifikante Verbindungen zwischen den Komponenten von semantischen Modellen in Komposita mit der Konstruktion V + N (χ^2; K)

LSU des Grundwortes / LSU des Bestimmungswortes	Person	Somatismen	Pflanzen	Stoffe und Materialien	Raum und Ort	Gebäude und Bauten	Gegenstände und Instrumente	Tätigkeit, Aktion	Seelische Sphäre	Sprache und Rede	Zeit, Alter	Kennwerte und Eigenschaften der Gegenstände	Sammelbezeichnungen von Menschen, Organisationen	Abstrakte Begriffe
Bewegung	5,58 0,08			0,09	0,96		0,52							
Geräusche und Kommunikation										29,63 0,19	6,09 0,09			
Verhalten													6,98 0,09	
Bearbeitung des Gegenstands	0,05		8,41 0,10	14,32 0,13	0,32		3,43							2,27
Physiologische Handlung						6,57 0,09			18,48 0,15					
Übergabe und Erhalten	1,74				1,09								24,23 0,17	
Physischer Einfluss auf ein Objekt		3,48					3,37	8,15 0,10				2,75		
Zustand					2,87		1,11							
Sensorische Wahrnehmung								18,12 0,15						
Geistige Arbeit												5,53 0,08		
Dasein							3,47							
Lage					2,45	41,41 0,23								

Tabelle G

Gebrauch und Rangverteilung von LSU der substantivischen Grundwörter
in den Komposita mit den Konstruktionen N + N, A + N, V + N

Wortbildungsmodell LSU des Grundworts	N + N	Rang	A + N	Rang	V + N	Rang
1. Person	689	33	73	34	45	31
2. Tiere	69	6	13	15	15	16
3. Somatismen	234	24	43	31	26	24
4. Attribute des Menschen	120	12	13	15	18	18
5. Pflanzen	209	23	28	26,5	31	27
6. Stoffe und Materialien	196	22	28	26,5	39	30
7. Raum und Ort	564	32	52	33	64	33
8. Gebäude und Bauten	404	29	26	23,5	51	32
9. Gegenstände und Instrumente	805	34	45	32	148	34
10. Essen und Getränke	129	14	10	10	12	13
11. Anzahl, Maßeinheiten	248	25	26	23,5	10	11
12. Bewegung	164	18	14	17	17	17
13. Tätigkeit, Aktion	320	28	34	29	37	29
14. Dasein	181	20	21	21	7	9
15. Possessorische Sphäre	60	3	8	7	2	2,5
16. Mentale Sphäre	136	15,5	12	13	19	19
17. Wahrnehmung	24	2	8	7	3	4,5
18. Seelische Sphäre	122	13	13	15	21	20,5
19. Sprache und Rede	291	27	25	22	30	26
20. Physiologische Sphäre	61	4	3	1,5	3	4,5
21. Verhalten und Handlungen	421	30	27	25	24	23
22. Eigenschaften des Menschen	64	5	11	11,5	13	15
23. Naturerscheinungen und Zustände	85	7	7	4,5	9	10
24. Physikalische Eigenschaften	95	10	7	4,5	12	13
25. Zeit, Alter	192	21	11	11,5	29	25
26. Kennwerte und Eigenschaften der Gegenstände	483	31	36	30	36	28
27. Veranstaltung, Spiel	113	11	9	9	6	7,5
28. Eigennamen	5	1	3	1,5	0	1
29. Staat, seine Attribute	136	15,5	18	19,5	2	2,5
30. Dokumente, Geld	88	8	6	3	6	7,5
31. Termini	92	9	8	7	5	6
32. Sammelbezeichnungen von Menschen, Organisationen	288	26	30	28	22	22
33. Abstrakte Begriffe	177	19	17	18	21	20,5
34. Wissenschaft, Kultur, Traditionen	141	17	18	19,5	12	13
Σ	7406		703		795	

Glottometrics 19, 2009, 42-48

Parts-of-speech diversification in Italian texts

Arjuna Tuzzi[1]*, Padua*
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the present article some characteristics of the rank-frequency distribution of parts-of-speech in the End-of-year speeches of Italian presidents are scrutinized. The result is compared with some other languages.

Keywords: Diversification, h-point, arc length, rank-frequency

In the endeavour to show that the Zipfian diversification is a law-like process whose indicator is different for different phenomena and does not depend or only secondarily depends on language, several researchers tried to find the properties of diversification (cf. Popescu, Altmann 2008; Fan, Altmann 2008; Fan, Popescu, Altmann 2008; Popescu, Kelih, Best, Altmann 2009; Laufer, Nemcová 2009; Sanada, Altmann 2009; Nemcová, Popescu, Altmann 2009; Popescu, Mačutek, Altmann 2009). Some distribution models of individual diversification phenomena were studied already in Rothe (1991a) who himself presented the wide scope of grammatical phenomena (Rothe 1991b). The linguistic background can be found in Zipf (1948) or in Köhler (1991).

In this contribution we scrutinize the diversification of parts-of-speech in a quasi-homogeneous collection of texts, namely in the End of Year speeches of Italian presidents beginning with Einaudi in 1949 and ending with Napolitano in 2008. The corpus is composed of sixty addresses, delivered by ten Presidents: Luigi Einaudi, Giovanni Gronchi, Antonio Segni, Giuseppe Saragat, Giovanni Leone, Sandro Pertini, Francesco Cossiga, Oscar Luigi Scalfaro, Carlo Azeglio Ciampi and Giorgio Napolitano. In Italy the duration of presidential term is seven years and each President usually delivered seven addresses (with three exceptions: Einaudi delivered the first address during the second year of his office, Segni resigned from his position after two years, Napolitano is the present President and is on his third year). The speeches are of different size, from 131 words with Einaudi in 1952 to 4916 with Scalfaro in 1995.

For the identification of parts-of-speech the state-of-the-art software tools currently available for the Italian language do not allow the full and correct lemmatization through a fully automated process. Lemmatization was conducted on the corpus through a partly manual (i.e.: disambiguating numerous lemmas by checking the context of occurrence) and partly automatic process (by means of the software Taltac2: www.taltac.it). The lemmatization process associated each token with a pair including a lemma and a grammatical category. In some cases the same token leads to different lemmas (as is the case with *potere*/power, noun or verb), in other cases different tokens are associated to the same pair (e.g. in the case of *nazione*/nation and *nazioni*/nations, which are both associated with the lemma *nazione* and category noun). The second effect often prevails and the number of types decreases. Owing to the wide range of contingent variations (masculine, feminine and plural forms, six different

[1] Address correspondence to: arjuna.tuzzi@unipd.it

persons, verb conjugations, clitic pronouns, etc.) in Italian the transition from a token-form vocabulary to a lemma-vocabulary may halve the number of different types.

For this study the following classes have been distinguished:

adjective
adverb
article
conjunction
interjection
noun
numeral
preposition
pronoun
proper noun
verb

and the analysis is valid under the condition that this classification is valid. As is well known the establishing of parts-of-speech or word classes is a source of incessant controversy. Our aim is to state the rank-frequency of theses classes in individual texts and show that all texts have something in common, namely a diversification constant which does not differ in different languages but distinguishes parts-of-speech from other linguistic phenomena. In a slightly metaphorical sense, it is a numerically expressible universal.

In order to characterize diversification we first state frequencies of parts-of-speech in individual texts, construct their rank-frequency distribution by ordering and compute the following properties of the rank-frequency sequence:

N = text size in terms of word form tokens.

R = inventory of parts-of-speech in individual texts which may vary from 8 to 11. It is at the same time the highest rank of the distribution.

$f(1)$ = the highest frequency in the distribution.

h = the h-point which is a fixed point of the rank frequency distribution computed in the form

$$h = \begin{cases} r & \text{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{if there is no } r = f_r \end{cases} \tag{1}$$

where r is the rank, f_r is the frequency at rank r. If there is a rank which is equal to its frequence, then $r = f_r$; if there is no such value, we take such (possibly neighbouring) values that $f_1 > r$ and $f_2 < r + 1$. If $r_2 = r_1 + 1$, the formula can be simplified. If $f(R) > R$, one must transform the whole ranked sequence in $f^*(r) = f(r) - f(R) + 1$. Consequently, in the following Table 1 and Table 2 the h-point and the text length N will be computed from this generalized $f^*(r)$ distribution. The importance of the h-point has been shown in many publications (cf. e.g. Popescu et al. 2009).

L = arc length computed as the sum of Euclidian distances between individual ordered frequencies, i.e. as

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

L_{max} = the maximal arc length which for rank-frequency distributions yields

(3) $Lmax = R - 1 + f(1) - 1$

and is necessary for stating the relative arc length.

Having all these numbers for each text we characterize the diversification by two indicators, namely

(4) $$p = \frac{L_{\max} - L}{h - 1}$$

which does not depend on text size, and

(5) $$q = \frac{L_{\max} - L}{N^{1/2}}$$

which can be constructed on the basis of the association of the *h*-point with text size *N*.

For the sake of illustration let us take the text of Einaudi 1949 in which we find the ordered rank-frequencies of parts-of-speech in the form

r	1	2	3	4	5	6	7	8	9
f(r)	41	37	33	30	17	15	14	6	1

Here $R = 9$, text size N is 194, the greatest frequency is $f(1) = 41$, the *h*-point lies evidently between 7 and 8 and can be computed according to (1) as

$$h = \frac{14(8) - 7(6)}{8 - 7 + 14 - 6} = 7.78.$$

Further, the arc length follows from

$$L = [(41 - 37)^2 + 1]^{1/2} + [(37 - 33)^2 + 1]^{1/2} + \ldots + [(6 - 1)^2 + 1]^{1/2} = 41.259.$$

The maximal arc length is simply

$$L_{max} = 9 - 1 + 41 - 1 = 48.$$

Using these numbers we obtain

$p = 0.994$
$q = 0.484.$

It is to be noted that if $f(R) > R$, one must transform all frequencies to obtain a new sequence $f^*(r) = f(r) - f(R) + 1$

Computing all these values for all texts we obtain the results in Table 1.

Table 1
Characteristics of parts-of-speech diversification in Italian End-of-Year Addresses

Text	Parts-of-speech frequency	*N*	*R*	*f**(1)	*h*	*L*	L_{max}	*p*	*q*
1949Einaudi	41;37;33;30;17;15;14;6;1	194	9	41	7,78	41,259	48	0,994	0,484
1950Einaudi	42;36;20;15;15;9;8;4;1	150	9	42	7,20	42,995	49	0,968	0,490
1951Einaudi	50,41,40,34,21,18,15,11	150	8	40	6,50	40,038	46	1,084	0,487
1952Einaudi	46,35,28,27,13,12,11,7	131	8	40	6,00	40,518	46	1,096	0,479
1953Einaudi	47,42,34,24,15,12,9,6,1	190	9	47	7,50	46,852	54	1,100	0,519
1954Einaudi	57,54,43,36,20,18,17,14,1	260	9	57	8,43	57,161	64	0,920	0,424
1955Gronchi	83,78,64,51,31,30,29,21,1	388	9	83	8,62	83,114	90	0,904	0,350
1956Gronchi	180,121,88,87,71,58,34,26	465	8	155	7,22	154,591	161	1,030	0,297
1957Gronchi	267,241,170,126,93,84,79,59,6,5	1090	10	263	8,87	262,656	271	1,060	0,253
1958Gronchi	201,162,131,127,82,74,63,42,3,1	886	10	201	8,85	200,544	209	1,077	0,284
1959Gronchi	181,135,92,80,72,71,36,29,1	697	9	181	8,72	180,644	188	0,953	0,279
1960Gronchi	196,161,112,106,78,63,45,38,3,2	794	10	195	8,80	194,686	203	1,066	0,295
1961Gronchi	304,244,184,162,111,105,75,65,2	1243	9	303	8,87	302,206	310	0,990	0,221
1962Segni	196,147,120,83,73,54,36,29	514	8	168	7,12	167,217	174	1,108	0,299
1963Segni	257,219,170,131,93,68,52,45,14,8	987	10	250	8,94	249,270	258	1,099	0,278
1964Saragat	102,85,84,64,42,40,28,17,3	447	9	100	8,47	99,850	107	0,957	0,338
1965Saragat	267,211,141,138,85,79,78,45,6,3	1033	10	265	8,87	264,875	273	1,032	0,253
1966Saragat	324,239,185,144,109,75,66,50,5,2	1189	10	323	8,89	322,316	331	1,101	0,252
1967Saragat	263,207,167,145,96,64,59,36,14,3,2	1045	11	262	9,33	261,673	271	1,120	0,289
1968Saragat	304,243,176,134,95,86,70,56,8,2	1164	10	303	8,74	302,256	311	1,130	0,256
1969Saragat	394,284,232,222,165,103,99,72,8,3,2	1573	11	393	8,97	392,744	402	1,161	0,233
1970Saragat	490,389,272,257,186,113,112,86,17,5,2	1918	11	489	9,54	488,701	498	1,089	0,212
1971Leone	70,51,37,35,30,17,11,6,3,2	252	10	69	7,50	69,194	77	1,201	0,492
1972Leone	182,149,134,111,69,45,45,24,5,3	747	10	180	8,70	180,389	188	0,988	0,278
1973Leone	298,232,205,174,103,97,76,63,1,1	1250	10	298	8,87	298,202	306	0,991	0,221
1974Leone	197,141,139,120,66,59,42,35,1,1	801	10	197	8,77	197,467	205	0,970	0,266
1975Leone	312,244,200,191,122,97,91,69,2	1319	9	311	8,88	310,214	318	0,988	0,214
1976Leone	321,239,211,196,113,112,97,73,3,1	1366	10	321	8,92	320,775	329	1,039	0,223
1977Leone	358,270,262,216,142,122,115,113,4,2	1594	10	357	8,94	356,658	365	1,051	0,209
1978Pertini	332,283,248,156,130,125,106,86,17,10	1403	10	323	8,91	322,278	331	1,103	0,233
1979Pertini	499,442,345,279,219,201,184,115,8,8,2	2291	11	498	8,84	498,182	507	1,125	0,184
1980Pertini	316,244,228,164,121,104,101,61,10,9,2	1349	11	315	9,00	314,757	324	1,155	0,252
1981Pertini	571,571,377,331,261,231,227,196,38,14,1	2818	11	571	10,29	571,239	580	0,943	0,165
1982Pertini	509,495,332,322,233,202,172,139,62,19,2	2476	11	508	10,44	507,190	517	1,039	0,197
1983Pertini	786,760,510,452,360,308,275,206,55,33,3	3726	11	784	10,68	783,110	793	1,022	0,162
1984Pertini	302,269,197,163,129,97,95,51,20,17	1180	10	286	8,84	285,493	294	1,085	0,248
1985Cossiga	612,427,404,289,207,192,120,93,10,3,2	2384	11	611	9,00	610,585	620	1,177	0,193
1986Cossiga	321,232,215,187,130,106,79,77,1,1	1349	10	321	8,90	321,344	329	0,969	0,208
1987Cossiga	501,414,349,248,184,163,107,103,11,11	1991	10	491	8,91	491,187	499	0,988	0,175
1988Cossiga	557,467,369,311,199,183,146,134,14,5	2345	10	553	9,10	552,170	561	1,090	0,182

1989Cossiga	441,399,302,231,154,145,102,101,31,6	1862	10	436	9,65	435,539	444	0,978	0,196
1990Cossiga	800,646,534,396,305,277,173,163,35,18	3177	10	783	9,50	782,123	791	1,044	0,157
1991Cossiga	95,71,64,57,48,29,26,22,4,2	408	10	94	8,68	93,794	102	1,069	0,406
1992Scalfaro	656,472,435,360,250,231,208,151,4,3,2	2761	11	655	8,96	654,916	664	1,141	0,173
1993Scalfaro	684,501,469,387,247,236,218,168,22,8,1	2941	11	684	9,86	683,221	693	1,104	0,180
1994Scalfaro	866,633,590,482,284,267,248,207,15,12,1	3605	11	866	10,17	865,299	875	1,058	0,162
1995Scalfaro	994,741,682,523,357,332,290,246,38,22,3	4206	11	992	10,50	991,120	1001	1,040	0,152
1996Scalfaro	535,348,326,313,183,128,115,110,16,11	1985	10	525	8,97	524,318	533	1,089	0,195
1997Scalfaro	1113,1048,712,522,429,397,368,329,54, 33,10	4916	11	1104	10,58	1103,110	1113	1,032	0,141
1998Scalfaro	972,775,577,415,399,289,254,251,35,23,5	3951	11	968	10,47	967,292	977	1,025	0,154
1999Ciampi	504,347,291,278,206,110,89,82,24,9,1	1941	11	504	9,94	503,262	513	1,089	0,221
2000Ciampi	432,338,291,273,168,124,95,88,23,12	1734	11	421	9,25	420,201	430	1,188	0,235
2001Ciampi	549,395,338,262,224,109,96,89,18,15,2	2086	11	548	10,29	547,353	557	1,038	0,211
2002Ciampi	556,389,312,304,209,132,112,98,10,7	2069	10	550	8,94	549,312	558	1,094	0,191
2003Ciampi	408,297,231,214,142,112,79,75,4,2,1	1565	11	408	8,93	407,861	417	1,153	0,231
2004Ciampi	455,353,268,265,147,111,93,88,19,8	1737	10	448	9,25	447,371	456	1,046	0,207
2005Ciampi	290,235,181,166,114,89,55,40,12,10,1	1193	11	290	10,00	289,439	299	1,062	0,277
2006Napolitano	502,377,356,286,191,169,159,146,10,7,1	2204	11	502	9,25	501,400	511	1,164	0,204
2007Napolitano	419,352,274,242,144,123,115,104,13,5,3	1772	11	417	9,22	416,470	426	1,159	0,226
2008Napolitano	409,328,281,220,135,127,120,86,4,2,1	1713	11	409	8,94	408,835	418	1,154	0,221

Different studies could be made having all these numbers (c.f. e.g. Cortelazzo, Tuzzi 2007; Pauli, Tuzzi 2009) but we are interested here only in diversification. The indicators are shown in the last two columns of Table 1. The mean of all p which do not depend on N is $\overline{p} = 1.062$ with $s_p = 0.070$, the mean of all q is $\overline{q} = 0.259$ with $s_q = 0.097$. Comparing these numbers with results on parts-of-speech from other languages we obtain the results presented in Table 2 where the sources are given in the following brackets: German 1, Latin, and Chinese (Schweers, Zhu 1991), German 2 (Best 1994), German SMS (Laufer, Nemcová 2009), Portuguese and Brazilian Portuguese (Ziegler 1998, 2001), Polish (Sambor 1989).

Table 2
Characteristics of parts-of-speech diversification in other languages

Text	Parts-of-speech frequency	N	R	$f^*(1)$	h	L	L_{max}	p	q
German 1	192, 161, 153, 112, 111, 104, 97,70	448	8	123	8	123	129	0.8889	0.2835
German 2	2032, 1939, 1532, 1338, 1179, 974, 914, 761	4589	8	1272	8	1271	1278	1.0574	0.1033
German SMS	2815, 2550, 2416,1606, 1459, 767, 541, 175	10937	8	2641	8	2640	2647	1.0574	0.0669
Portuguese	2586, 1607, 949, 819, 776, 680, 478, 440, 352	5528	9	2235	9	2234	2242	1.0582	0.1076
Brazilian Portuguese	2930, 2265, 1743, 1708, 1602, 1040, 936, 394	9474	8	2537	8	2536	2543	1.0014	0.0719
Latin	347, 173, 142, 98, 93, 59, 40, 39, 9	928	9	339	9	339	346	0.9044	0.2298
Polish	144188, 79995, 71988, 56812, 33605, 31833, 21428, 18757, 8076, 650	460842	10	143539	10	143538	143547	1.0000	0.0133
Chinese	247, 228, 140, 133, 107, 81, 55, 27	810	8	221	8	220	227	1.0072	0.2460

Conclusions

In Italian *p* is slightly greater than in other languages but evidently this quantity is a general language constant. The fact that in Polish furnishing the longest text(s), it is exactly 1.0000 is symptomatic and very conspicuous but until further data are not at our disposal nothing can be followed. In Italian, one can observe a decrease of the variance of *p* with increasing *N* but this is only a preliminary statement. In Figure 1, the independence of *p* itself on *N* is evident, hence for interlinguistic comparison *q* is more appropriate. The indicator *q* heavily depends on *N* hence using it for comparative purposes one must take *N* into account (in tests for difference, N is part of the variance).

Using the results from previous investigations, we can see that *p* should be analyzed in many languages to see its general status, and *q* can be used for typological and stylistic purposes.

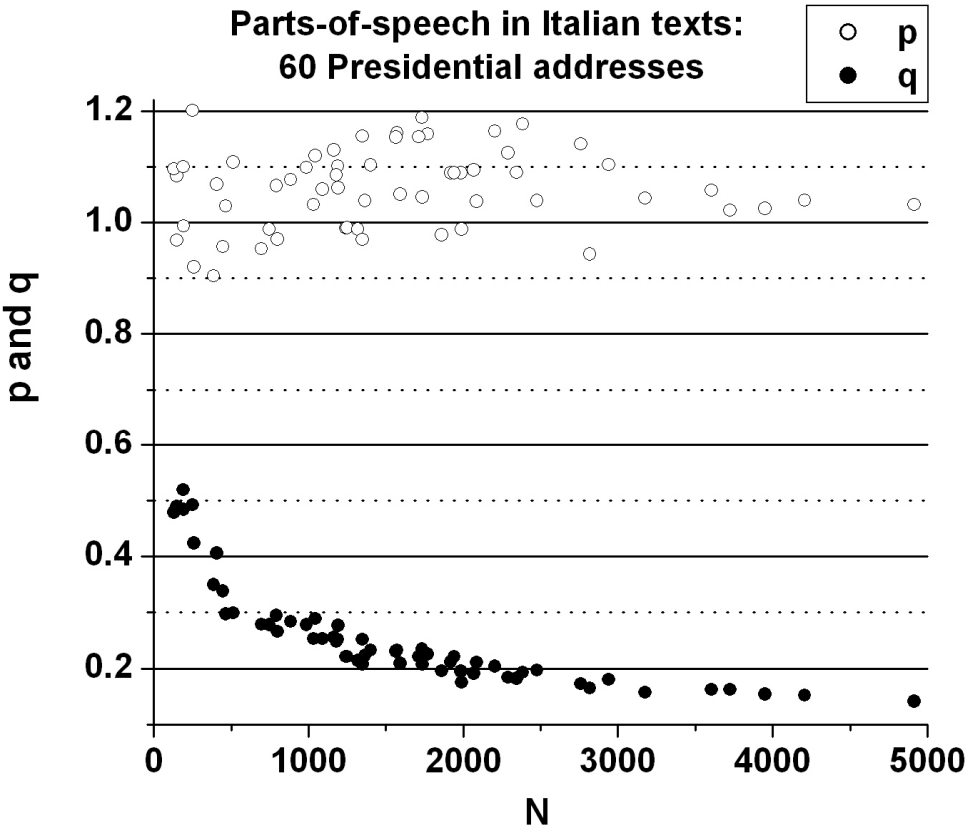


Figure 1. Indicators *p* and *q* for parts-of-speech in 60 Italian texts

References

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

Best, K.-H. (2008). Diversifikation des Phonems /r/ im Deutschen**.** *Glottometrics 18, 2009, 26-31.*

Cortelazzo, M.A., Tuzzi, A. (eds.) (2007). *Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica.* Venezia: Marsilio.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 66-78.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 79-86.*

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe (1991a): *47-55*.

Laufer, J., Nemcová, E. (2008). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 2009, 13-25*

Nemcová, E., Popescu,I.-I., Altmann, G. (2009). Word associations in French**.** In: *Festschrift R. Grotjahn (submitted).*

Pauli, F., Tuzzi, A. (2009). The End of Year Addresses of the Presidents of the Italian Republic (1948-2006): similarities and differences. *Glottometrics 18, 2009, 40-52*

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottomerics 17, 94-108.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics* 18, 32-39.

Popescu, I.-I., Mačutek, J., Altmann, **G.** (2009). *New aspects of word frequencies*. (submitted).

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

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

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

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics 19 (submitted).*

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

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

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

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

Glottometrics 19, 2009, 49-69

A word length regularity and its genesis

Mats Eeg-Olofsson, Lund[1]

Abstract. Data from several typologically and genetically distinct languages show that the distribution of length in letters or phonemes of word types fits the Good distribution, the discrete analogue of the Gamma distribution. It is suggested that this is an equilibrium distribution, resulting from long oral tradition of language between successive generations. Length redistribution can take place when juxtaposed words in speech are re-interpreted by second generation listeners. Simulations show that a certain kind of random shifting of word boundaries results in a lexicon whose word lengths are also well fitted by the Good distribution. The fit remains close even if a limited amount of compounding is introduced into the simulations.

Keywords: word length, Good distribution, English, German, Hungarian, Italian, Russian, Swedish

1. Introduction

Knowledge about word length regularity is important in stylistics as well as in practical language engineering tasks such as cryptography, text alignment and text segmentation. Word length regularity is also interesting as a possible universal characterizing human language.

The words to be measured here are word types, the different words making up the vocabulary of a text or a text corpus. Such a vocabulary is taken to represent the vocabulary of the underlying language system.

Word length can be measured in terms of various units, such as morphemes, syllables, phonemes, and letters. The following data are mostly measured in terms of letters, in some cases in phonemes.

While a considerable amount of work has been done on word length measured in syllables (see e.g. the bibliography of Best 2009), there is relatively little research about word length counted in letters or phonemes. Some exceptions are Herdan (1958), who argues for a description using the lognormal distribution, and Alekseev (1998).

Graphs describing the length distribution of the vocabulary of some well-known languages have a characteristic shape, as shown by Figure 1 for American English represented by the Brown Corpus (Kučera & Francis 1967) and by Figure 2 for Swedish according to NFO1, the Frequency Dictionary of Present-Day Swedish (Allén 1970). The empirical distributions have been fitted by the Good distribution (see below), as described in the tables. In the figures circles denote observed frequencies, while crosses denote theoretical frequencies, which have been rounded to the nearest integer.

[1] Address correspondence to: Mats Eeg-Olofsson: matseeg@gmail.com.

Table 1
Word length in the Brown Corpus fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	36	4	20	56	36
2	277	160	21	46	19
3	1193	1000	22	19	10
4	3025	2773	23	10	5
5	4619	4928	24	6	3
6	6470	6606	25	9	1
7	7356	7290	26	6	1
8	6999	6978	27	7	0
9	6073	5991	28	3	0
10	4823	4721	29	2	0
11	3436	3471	30	2	0
12	2335	2410	32	1	0
13	1443	1594	33	1	0
14	881	1013	34	1	0
15	534	621	36	1	0
16	331	369	37	1	0
17	194	214	38	1	0
18	120	121	41	1	0
19	87	67	44	1	0
η = -6.9033, q = 0.3808, R^2 = 0.9984					

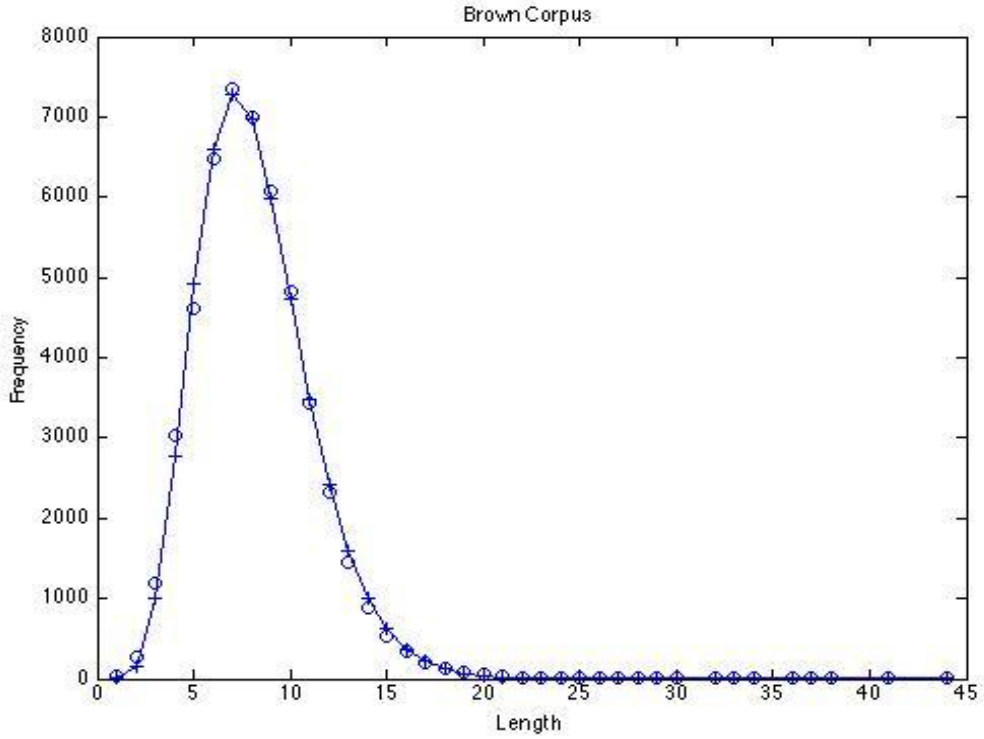


Figure 1. Length distribution of word types in the Brown Corpus

Table 2
Word length in Nusvensk frekvensordbok fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	45	21	19	1354	1188
2	378	367	20	942	849
3	1510	1540	21	722	599
4	3529	3583	22	442	418
5	5884	6039	23	279	289
6	8409	8302	24	155	197
7	9848	9915	25	103	134
8	10640	10683	26	66	90
9	10814	10639	27	33	60
10	10082	9958	28	24	40
11	8856	8864	29	11	26
12	7401	7570	30	4	17
13	5973	6244	31	7	11
14	4757	5000	33	3	5
15	3862	3902	34	1	3
16	3045	2979	59	1	0
17	2392	2230	61	1	0
18	1843	1641			
η = -4.9948, q = 0.5530, R^2 = 0.9992					

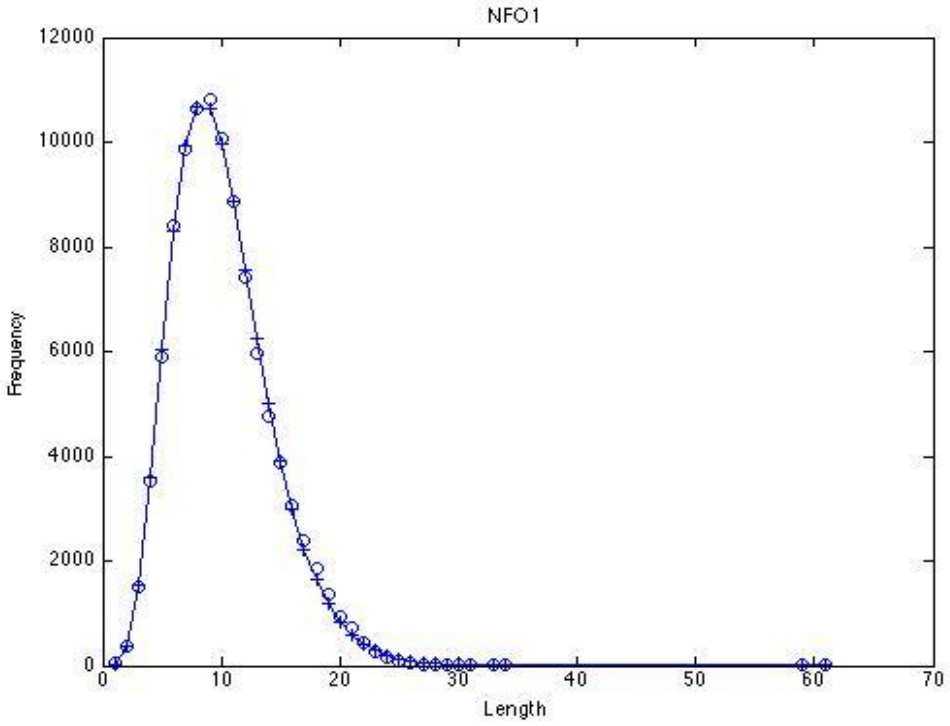


Figure 2. Length distribution of word types in Nusvensk frekvensordbok

Bengt Sigurd has suggested that such graphs can be described by functions defined by the formula

$$f(x) = C\, x^a\, b^x,$$

where $f(x)$ is the frequency of words having length x, a and b are parameters with positive values, and C is a normalizing constant. It should be emphasized here that the a and b parameters take on different values for different languages.

The power function x^a reflects the simple fact that there are more possible long words than short words, which naturally results in higher frequencies. The damping factor b^x, where b lies between 0 and 1, is motivated by the fact that long words are less economical, demanding more resources for production and understanding.

Sigurd et al. (2004) fitted the above formula to various linguistic length data. Similar results were found by Lupsa & Lupsa (2005) for base forms in Romanian and English dictionaries, and by Eeg-Olofsson (2008) for the vocabulary of various languages (English, Finnish, French, Sorbian, Swedish, and Turkish) as represented in the *Leipzig Corpora Collection*, using the curve-fitting software Regress+.

Naturally, this observed regularity calls for an explanation. For a relation between word length and lexical access (reaction) time, see Sigurd et al. (forthcoming).

The above formula is a notational variant of the formula describing the frequency function of the Gamma distribution

$$p(x) \;=\; c(\alpha,\theta)\; x^{\alpha-1}\; e^{-x/\theta},$$

The constant c is determined by the parameters α and θ, equalling $1/(\theta^{\alpha}\, \Gamma(\alpha))$, where Γ is the Gamma function. α and θ are called the shape and the scale parameter, respectively, of this distribution.

The discrete analogue of the Gamma distribution is the so-called Good distribution (Johnson et al. 2005), whose probability function is

$$p(x) \;=\; k\; x^{-\eta}\; q^{x},$$

where the length x takes on natural numbers 1, 2, 3, ..., the parameter q is a number between 0 and 1, and the normalizing constant k depends on the values of the η and q parameters, equalling $1/Li_{\eta}(q)$, where Li is the polylogarithm function (see e.g. Weisstein 2009). In Section 2 further empirical data from various languages are fitted by the Good distribution.

As a tool for describing the length of linguistic units, the Good distribution can be derived in various ways. It is the stationary distribution of various birth-death processes, as noted by Kulasekera and Tonkyn (1992). A particular choice is

$$\lambda_x \;=\; \left(1 + x^{-1}\right)^{-\eta}$$

for the birth intensities and

$$\mu_{x+1} \;=\; q^{-1}$$

for the death intensities, $x = 1, 2, \ldots$

The birth intensities can be interpreted as a decreasing lengthening tendency, which may be caused by the need to expand the vocabulary. The death intensities represent a general tendency towards parsimony.

The above line of reasoning is closely related to the unified approach of Wimmer and Altmann (2005). The Good distribution has a recurrence relation of the form $P(x+1) = g(x)P(x)$, where $g(x) = q(1+x^{-1})^{-\eta}$, whose factors correspond to the above intensities.

Section 3 describes yet another derivation, a stochastic process that makes less specific assumptions about the form of the functions that determine the dynamics of language behaviour. It is a kind of "exchange game", a lexical restructuring process, which can be viewed as a kind of cultural evolution.

Simulations reported in Section 4 indicate that the Good distribution is a good approximation of the equilibrium distribution related to the above-mentioned stochastic redistribution process.

2. Empirical data

Additional word length data from various published corpora or word lists have been fitted by the Good distribution. Orthographic data represent the languages German, Hungarian, Italian, and Russian. There are also some Swedish and English phonetic data. The computations use the Curve Fitting Toolbox of the MATLAB programming package. The polylogarithm values have been computed by means of the MATLAB polylog.m function made available by Willem Ottevanger (2009). The fit is very good, R^2 exceeding 0.99 in all cases.

2.1 Orthographic data

Care has been taken to make the words in the orthographic lists as close to a phonemic transcription as is permitted by conventional orthography. The units to be counted are basically graphic words, delimited by spaces and subsequently stripped of leading and trailing punctuation signs, such as parentheses and commas. Abbreviations containing punctuation marks have been excluded. Digital numerals, both pure *(1984)* and hybrid *(19th)* are also omitted. Word-internal apostrophes and other punctuation signs with similar function have been deleted from the word bodies prior to further processing. Only such words have been accepted as consist of sequences of letters, optionally separated by hyphens. Finally, all letters have been turned into lower case, no distinction being made between, e.g., *china* and *China*. Hyphens are ignored in the computation of the length of the selected words.

The German data are taken from the 100,000-sentence German sample of the Leipzig Corpora Collection. The original data contain multi-word units, which have been split up into the component graphic words. The results of fitting can be seen in Table 3 and Figure 3.

Table 3
Length of German word types fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	32	22	21	1613	1782
2	348	368	22	1142	1340
3	2088	1534	23	840	997
4	3746	3643	24	563	736
5	6959	6360	25	394	538
6	9596	9138	26	266	391
7	10911	11478	27	175	282
8	12251	13067	28	89	202
9	13401	13799	29	71	144
10	14120	13733	30	52	102
11	13673	13028	31	23	72
12	12467	11879	32	14	50
13	10661	10478	33	19	35
14	8857	8984	34	11	24
15	7514	7517	35	6	17
16	6186	6157	36	4	12
17	5009	4950	37	1	8
18	3889	3915	39	2	4
19	2894	3050	44	1	1
20	2304	2346			
η = -4.7705, q = 0.6021, R^2 = 0.9964					

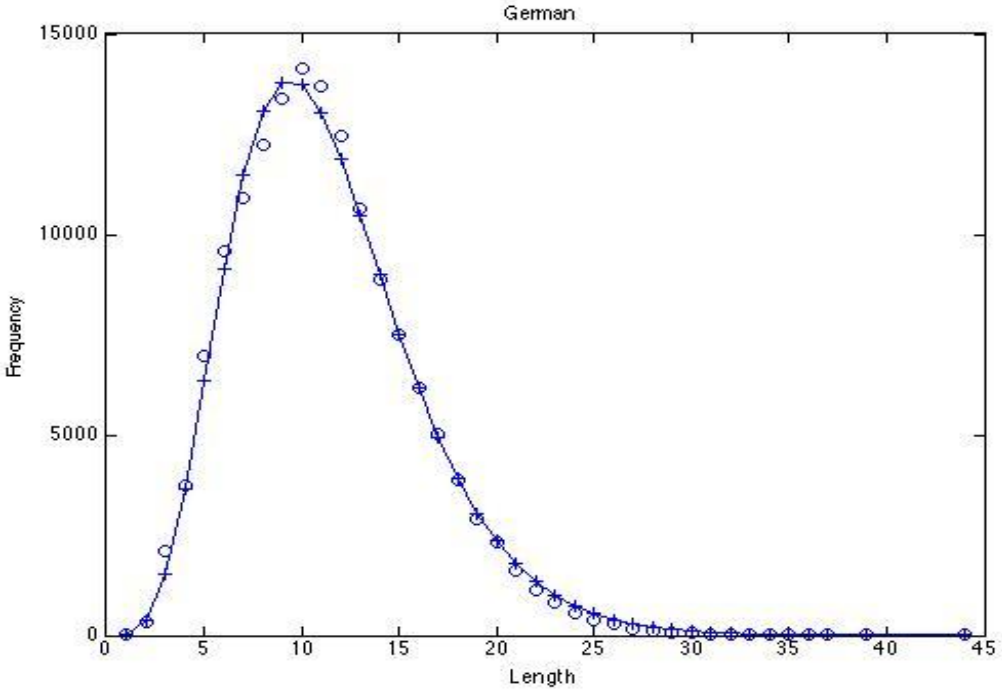


Figure 3. Length distribution of German word types in the Leipzig Corpora

The data for Hungarian are taken from the "best" 4% part of the published word list of the Hungarian webcorpus (Halácsy et al. 2004, Kornai et al. 2006). This relatively big corpus contains a large number of long list-like compounds, e.g. route descriptions like "Berlin-Paris-London". Figure 4 shows the length distribution of Hungarian words with length falling below 30 letters, excluding many such compounds. However, the goodness of fit exceeds 0.99 independently of whether longer words are included or not.

Table 4
Length of Hungarian word types fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	35	1	16	321355	322228
2	1050	173	17	245575	251653
3	13032	2941	18	179753	190546
4	40826	17445	19	128419	140349
5	91686	58151	20	88889	100847
6	162808	134629	21	59271	70861
7	245658	242366	22	39572	48792
8	352076	362947	23	25250	32981
9	445480	472218	24	16342	21921
10	534056	549879	25	10537	14345
11	584613	585352	26	6770	9254
12	587709	578607	27	4390	5892
13	552138	537436	28	2429	3705
14	488633	473428	29	150	2304
15	404730	398424			

η = -8.9294, q = 0.4545, R^2 = 0.9959

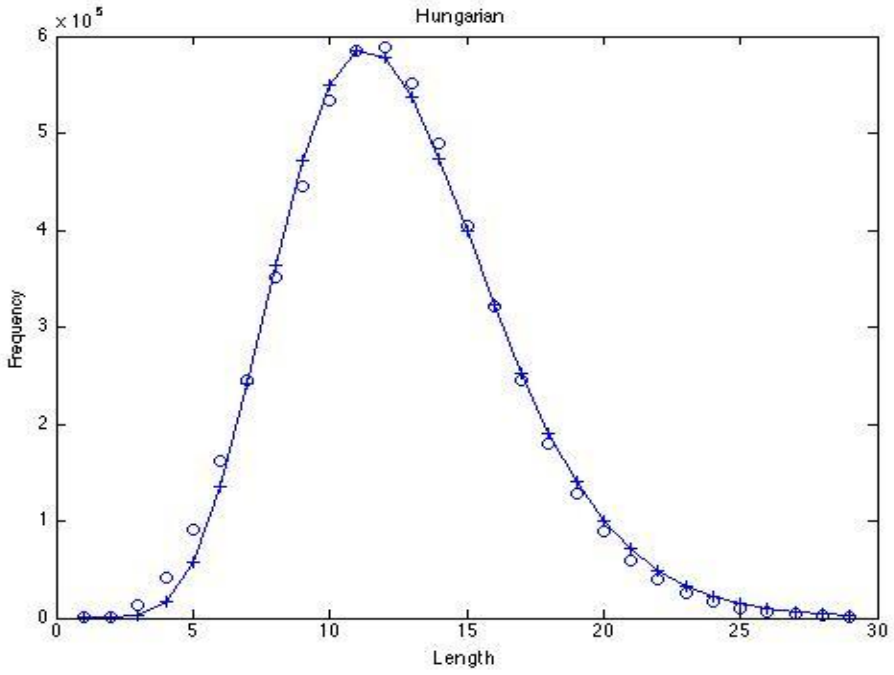


Figure 4. Length distribution of word types in the Hungarian webcorpus

Like the German data, the Italian data are taken from a 100,000-sentence sample in the Leipzig Corpora Collection.

Table 5
Length of Italian word types fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	30	0	17	390	475
2	254	35	18	232	256
3	1184	474	19	124	134
4	2850	2153	20	77	69
5	5700	5384	21	51	34
6	8261	9235	22	32	17
7	12188	12211	23	18	8
8	13086	13348	24	16	4
9	12884	12615	25	6	2
10	11164	10631	26	2	1
11	8164	8164	27	2	0
12	5369	5807	28	2	0
13	3524	3873	29	1	0
14	2167	2446	31	1	0
15	1196	1473	33	1	0
16	725	851			
η = -9.2355, q = 0.3185, R^2 = 0.9948					

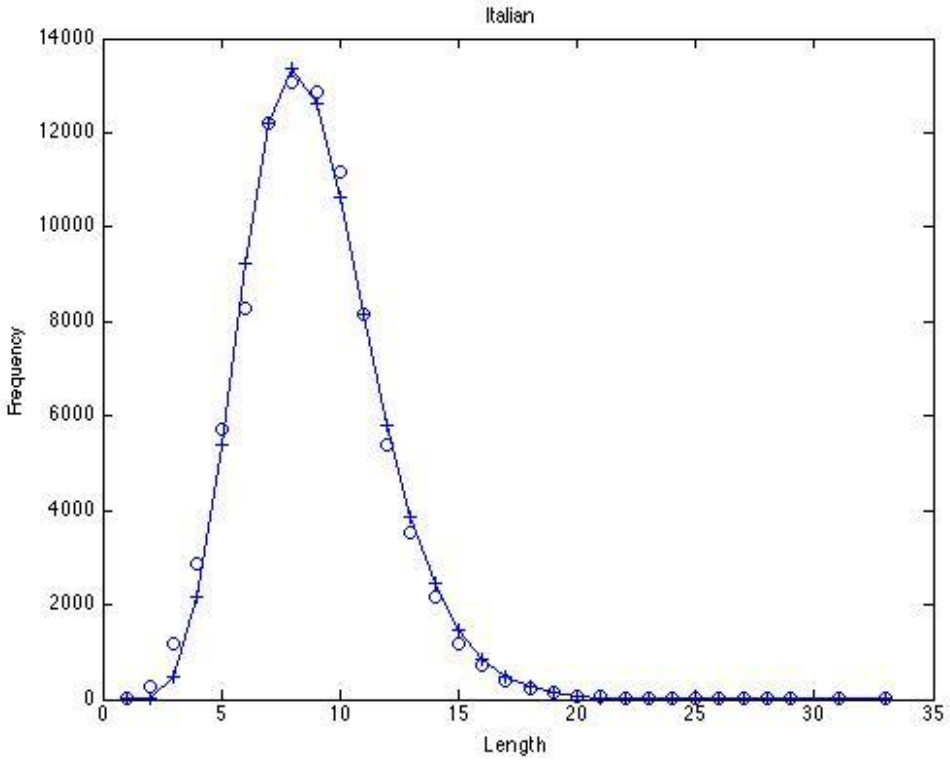


Figure 5. Length distribution of Italian word types in the Leipzig Corpora

Russian word length statistics in terms of Cyrillic letters have been computed from a machine-readable version of the Uppsala Russian Corpus (Lönngren 1993). The results are presented in Table 6 and Figure 6.

Table 6
Length of Russian word types fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	26	0	17	952	879
2	133	59	18	658	496
3	892	716	19	377	272
4	3034	3066	20	281	146
5	7293	7447	21	158	77
6	12314	12632	22	120	39
7	16951	16725	23	91	20
8	18388	18472	24	66	10
9	17501	17764	25	50	5
10	15293	15314	26	31	2
11	12236	12086	27	23	1
12	8749	8867	28	22	1
13	5626	6119	29	12	0
14	3967	4008	30	5	0
15	2431	2509	31	3	0
16	1562	1511			
η = -8.7933, q = 0.3414, R^2 = 0.9994					

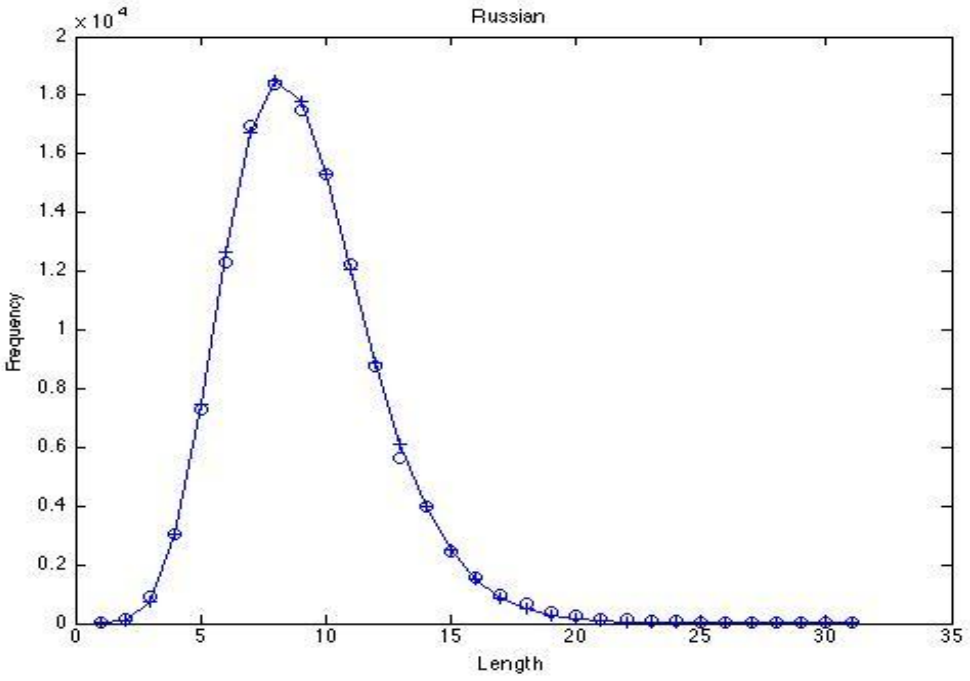


Figure 6. Length distribution of Russian word types in the Uppsala Corpus

2.2 Phonetic data

The *MRC Psycholinguistic Database* contains length data in phonemes for 32,549 different English words as presented in Table 7 and Figure 7.

Table 7
Length of word types in the MRC Database fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	32	15	11	1536	1479
2	276	347	12	862	942
3	1442	1484	13	450	575
4	3396	3138	14	206	339
5	4561	4511	15	110	194
6	4985	5079	16	42	108
7	4691	4832	17	9	59
8	4199	4064	18	3	31
9	3317	3110	19	3	16
10	2429	2210			
$\eta = -5.9810$, $q = 0.3784$, $R^2 = 0.9957$					

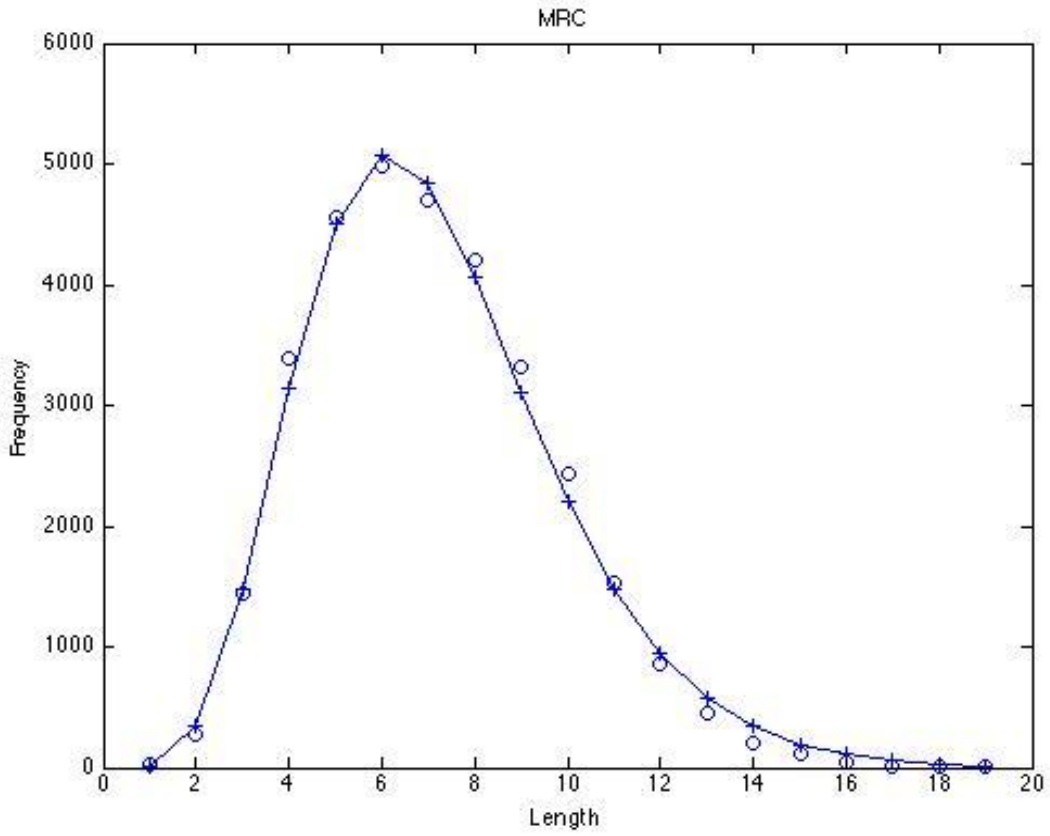


Figure 7. Length distribution in phonemes of word types in the MRC Database

Swedish phonetic data are taken from a machine-readable version, including inflected word forms, of a Swedish pronouncing dictionary, *Norstedts svenska uttalslexikon* (1997). The results are presented in Table 8 and Figure 8.

Table 8
Length of word types in Norstedts svenska uttalslexikon fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	37	0	17	31193	31992
2	160	46	18	23730	24034
3	1673	660	19	17886	17591
4	5260	3468	20	13071	12577
5	12854	10575	21	9345	8804
6	25509	22845	22	6323	6046
7	38627	38911	23	4091	4080
8	52388	55690	24	2567	2710
9	66540	69784	25	1494	1774
10	78487	78742	26	811	1145
11	83764	81624	27	434	730
12	81266	78889	28	221	460
13	73489	71891	29	90	287
14	61420	62313	30	42	177
15	49145	51729	31	19	108
16	39263	41358	32	4	66
η = -8.4394, q = 0.4638, R^2 = 0.9975					

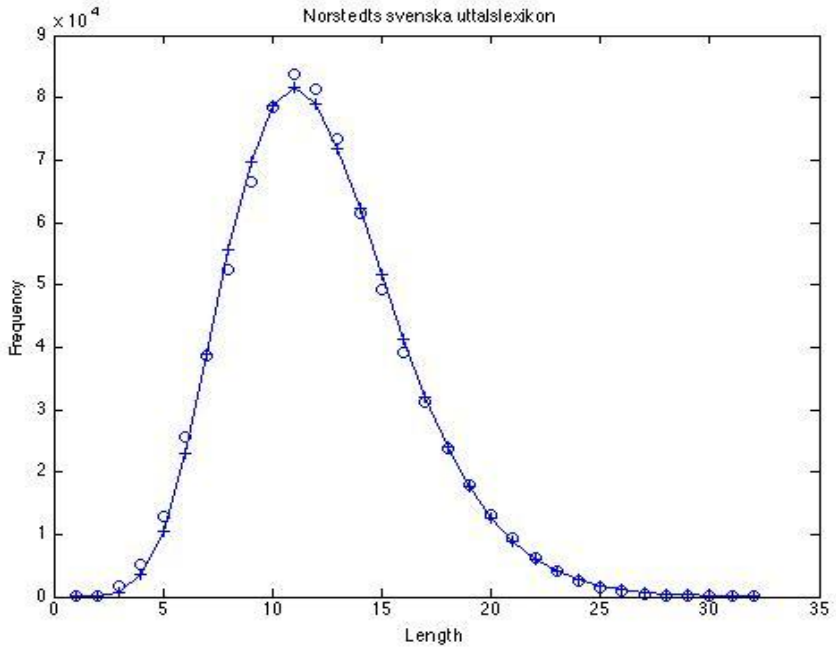


Figure 8. Length distribution in phonemes of word types in *Norstedts svenska uttalslexikon*

2.3 Survey

The new data visualized in Figures 3-8 are summarized in Table 9.

Table 9
Survey of results from Tables 3 to 8

Corpus	**Number of types**	**Mean type length**	**η**	**q**	**R^2**
German (Leipzig)	152,192	11.2	-4.77	0.6021	0.9964
Hungarian Webcorpus	5,633,232	12.4	-8.929	0.4545	0.9959
Italian (Leipzig)	89,701	8.8	-9.235	0.3185	0.9948
Russian (Uppsala)	129,245	9.2	-8.973	0.3414	0.9994
MRC Psycholinguistic Database	32,549	7.1	-5.981	0.3784	0.9957
Norstedts svenska uttalslexikon	781,203	12.2	-8.439	0.4638	0.9975

3. The redistribution process

Lexical redistribution of word length can take place any time two words (or other lexical units) are juxtaposed in speech. Obviously, the hearer and the speaker can interpret such a syntagm in different ways, by placing the word boundary differently. Communication still works because of the overall context.

A well-known example from historical linguistics is the so-called nasal shift in English (Barnes 1980) underlying the formation of the word *adder* from *nadder* by reinterpretation of *a nadder* as 'an adder'. A new, shorter designation of the same concept has been created. The same process in reverse direction transforms *ekename* into *nickname*. A similar process is believed (Wessén 1979, p. 122) to have created the Swedish pronoun *ni* (plural 'you') from an older *I* ('ye'). In this case a new, longer word has been created by restructuring of sequences such as *kunnen I* ('can ye') as 'kunne ni'. While some such changes can be detected in the relatively few languages that have written records, they are even more likely to take place when language is inherited by one generation from the preceding generation by oral tradition only.

The redistribution process proceeds in a large number of steps, each step involving two different words that are selected randomly from the vocabulary and juxtaposed. The word boundary is shifted randomly, which may result in the formation of two new word forms, which replace the original ones. Obviously inter-generation communication may break down if the changes are too radical. Therefore, it seems reasonable to assume that a certain proportion of the original word forms must be preserved in the new word forms. This can be quantified by a so-called saving propensity factor, denoted by the symbol λ, a number between 0 and 1.

With these specifications, the redistribution step can be described more formally in the following way:

Two different words out of a vocabulary of V words are selected randomly. The words, which turn out to be word number i and word number j, have the initial sizes x_i and x_j, respectively. In the redistribution they retain at least λx_i and λx_j, respectively, of their initial

size. The remaining $(1-\lambda)(x_i + x_j)$ is split randomly. Consequently, the new sizes x_i' and x_j' of the words after the interaction can be expressed as

$$x_i' = \lambda x_i + \varepsilon(1-\lambda)(x_i + x_j)$$

and

$$x_j' = \lambda x_j + (1-\varepsilon)(1-\lambda)(x_i + x_j)$$

where ε is a random number between 0 and 1, describing the random shifting.

It may be noted that long words (i.e. long designations of concepts) fluctuate more in this process than short words.

The above calculation must be modified slightly by rounding when the lengths must be natural numbers. In that case the process is simply a Markov chain, whose states correspond to compositions into V positive integers of the total length of the vocabulary. The equilibrium length distribution is closely related to the equilibrium state distribution of this Markov chain.

I have not found any analytical derivation of the equilibrium length distribution of this process, but the simulations reported in Section 4 show that it is well approximated by the Good distribution. For the corresponding process with continuously varying lengths the situation is similar. Numerical simulations by Patriarca et al. (2004) have produced data that are well fitted by the closely related Gamma distribution, but according to Chatterjee and Chakrabarti (2007) the exact form of the distribution is still to be found.

Appendix 1 contains code in the programming language Perl describing the redistribution process, for the convenience of those who wish to make simulations like those described in Section 4.

4. Simulations

4.1 Redistribution of length by boundary shifting

The above-mentioned redistribution process has been simulated on computer for various choices of the number of words, saving propensity, and initial size distribution. The simulations suggest that the process rapidly reaches a steady state that is well approximated by the Good distribution. It should be emphasized that the simulations are to be regarded as time-compressed views of real language change. In real linguistic communication, words are not likely to be restructured every time they are juxtaposed in speech.

Figure 9 shows a simulation with saving propensity 0.7 and 10,000 units, each having the initial size 8. The graphs show the distributions after 10,000, 20,000, 30,000, 40,000, and 50,000 steps. Apparently, the graphs coalesce into a curve describing a steady state distribution.

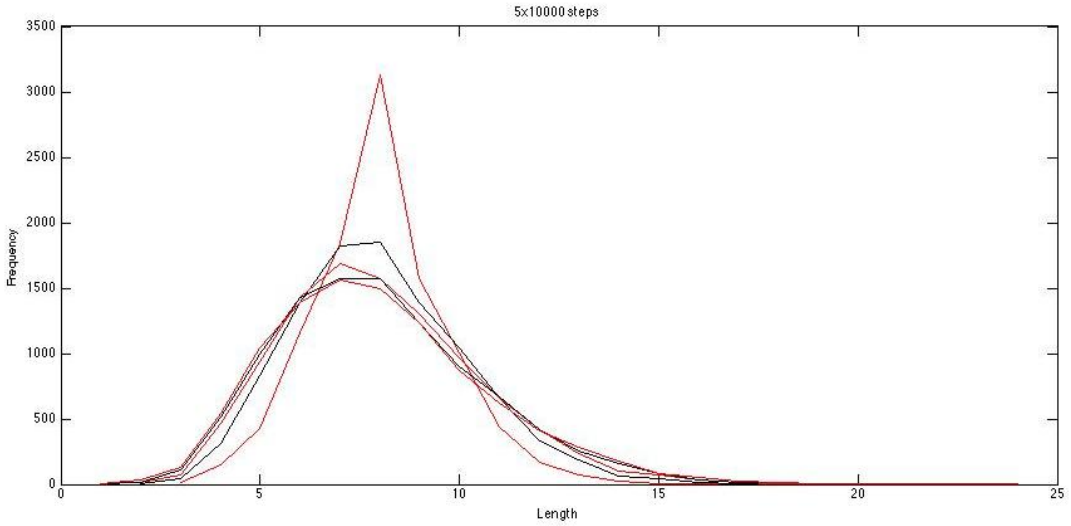


Figure 9. Simulation in five successive steps

Figure 10 shows five different simulations with the same parameters, each comprising one hundred million steps. Clearly, the end results are very close to each other. As shown in Table 10, the data pooled from these five simulations are also well fitted by the Good distribution.

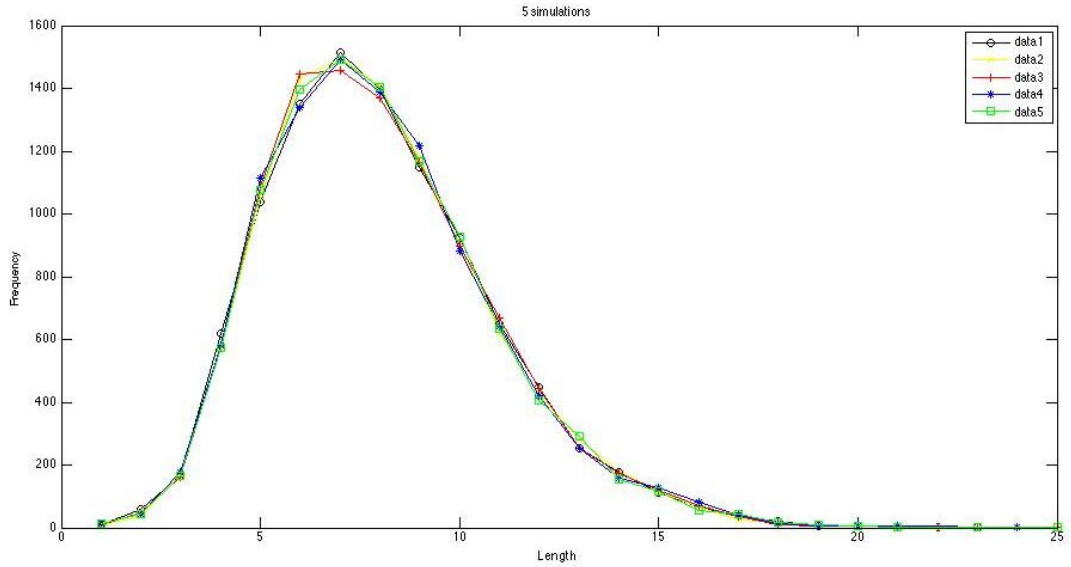


Figure 10. Five simulations with identical parameters

Table 10
Data pooled from five simulations fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
1	52	3	14	845	845
2	231	163	15	589	499
3	848	1058	16	340	286
4	2921	2962	17	179	159
5	5364	5241	18	79	87
6	6965	6937	19	35	46
7	7448	7514	20	22	24
8	6962	7029	21	15	12
9	5879	5879	22	5	6
10	4531	4501	23	4	3
11	3210	3208	24	1	2
12	2142	2155	25	3	1
13	1330	1378			
$\eta = -7.1133$, $q = 0.3618$, $R^2 = 0.9994$					

Other simulations have shown that equilibrium can be reached also when the initial distribution is widely different from a uniform one. One such simulation used an initial distribution where 9,999 units had size 1, whereas the remaining unit had size 70,001.

4.2 Boundary shifting and compounding

The vocabulary simulated in the above way is obviously a static one, with a fixed total size and a fixed number of units (senses). In order to accommodate some linguistic creativity, some simulations have also included a limited degree of "compounding". In a small proportion of the restructuring steps, the change consists in concatenation of the words rather than boundary shifting. Thus new words are created, increasing the number of senses and the total size of the lexicon. This is another restructuring device, which may be called compounding.

Some natural languages use compounding extensively, e.g. German and Swedish. Among other things, compounds are created as designations of new concepts, e.g. German *Kraftwagen* for 'car'. Compounding is sometimes also used to clarify or disambiguate opaque or ambiguous simplex words. A case in point is Chinese, which contains a large proportion of such compounds, each component of which is highly ambiguous because of excessive homophony. In an example such as *péngyou* ´friend´ the ambiguous components *péng* and *yǒu* disambiguate each other mutually.

Table 11 and Figure 11 show an extended simulation, where every 10,000th step is compounding rather than word boundary shifting. Thus, in 100,000,000 steps in all, about 10,000 new units are created in addition to the initial 10,000. The resulting length distribution is still very well fitted by the Good distribution, but of course the mean word length is greater than before.

Table 11
Simulation data with 0.01% compounding fitted by the Good distribution

Length	Observed	Theoretical	Length	Observed	Theoretical
3	2	8	25	357	361
4	30	36	26	290	289
5	104	103	27	250	229
6	218	221	28	184	180
7	401	390	29	135	141
8	587	598	30	129	109
9	833	823	31	82	83
10	1031	1039	32	66	64
11	1207	1225	33	40	48
12	1364	1365	34	29	36
13	1470	1448	35	30	27
14	1481	1476	36	19	20
15	1461	1452	37	13	15
16	1426	1386	38	11	11
17	1237	1287	39	10	8
18	1132	1168	40	6	6
19	1033	1037	41	8	4
20	918	903	43	4	2
21	808	773	44	1	2
22	637	652	46	1	1
23	537	542	54	1	0
24	438	445			
η = -6.8717, q = 0.6124, R^2 = 0.9992					

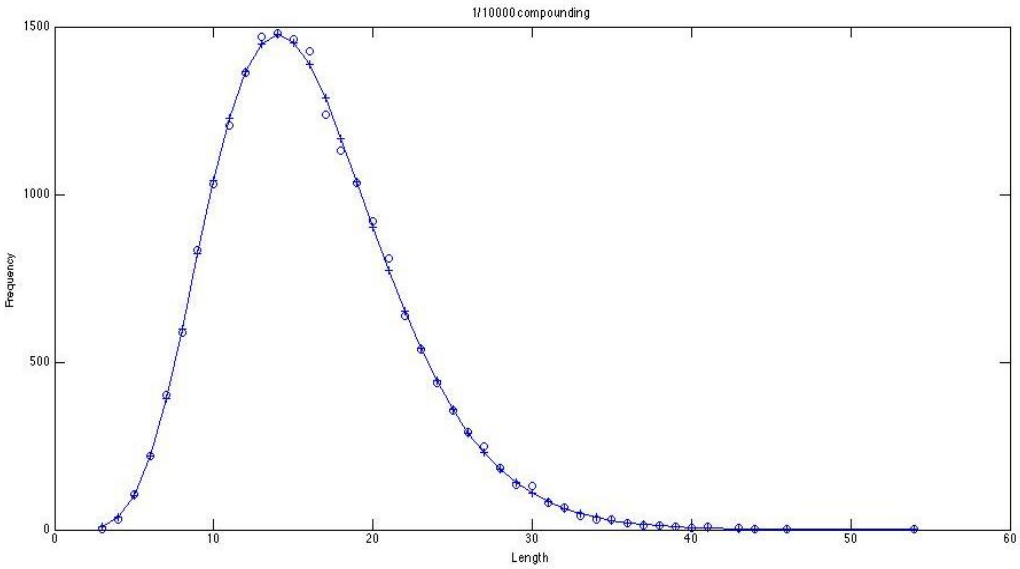


Figure 11. Simulation including 0.01% compounding

In Table 12 and Figure 12 the compounding rate has been increased from 0.0001 to 0.0003, resulting in a quadruple size vocabulary after 100,000,000 steps. Again, mean length has increased and some very long units have been created, but the distribution is still very close to the expected Good distribution.

Table 12
Simulation data with 0.03% compounding fitted by the Good distribution

Length	Obs.	Theor.	Length	Obs.	Theor.	Length	Obs.	Theor.
5	2	3	36	1197	1181	67	43	43
6	6	8	37	1099	1117	68	33	37
7	19	18	38	1067	1050	69	28	32
8	32	35	39	969	983	70	23	28
9	45	62	40	974	916	71	22	24
10	77	101	41	873	850	72	24	21
11	154	152	42	821	785	73	18	18
12	191	217	43	746	723	74	17	15
13	261	295	44	662	663	75	22	13
14	384	385	45	638	605	76	6	11
15	490	485	46	482	551	77	15	10
16	586	592	47	504	500	78	7	8
17	701	704	48	449	453	79	6	7
18	803	817	49	434	408	80	7	6
19	906	929	50	389	367	81	6	5
20	1057	1036	51	292	329	82	3	4
21	1148	1135	52	280	295	83	3	4
22	1234	1224	53	291	263	84	4	3
23	1388	1302	54	211	234	85	4	3
24	1349	1367	55	222	208	86	1	2
25	1413	1417	56	188	184	87	2	2
26	1480	1454	57	189	163	88	1	2
27	1474	1476	58	132	144	89	3	1
28	1487	1484	59	132	126	90	3	1
29	1515	1480	60	111	111	91	1	1
30	1437	1463	61	90	97	100	1	0
31	1440	1435	62	94	85	114	2	0
32	1320	1398	63	89	74	122	1	0
33	1328	1353	64	75	65			
34	1272	1301	65	57	57			
35	1212	1243	66	40	49			
$\eta = -6.9074$, $q = 0.7823$, $R^2 = 0.9983$								

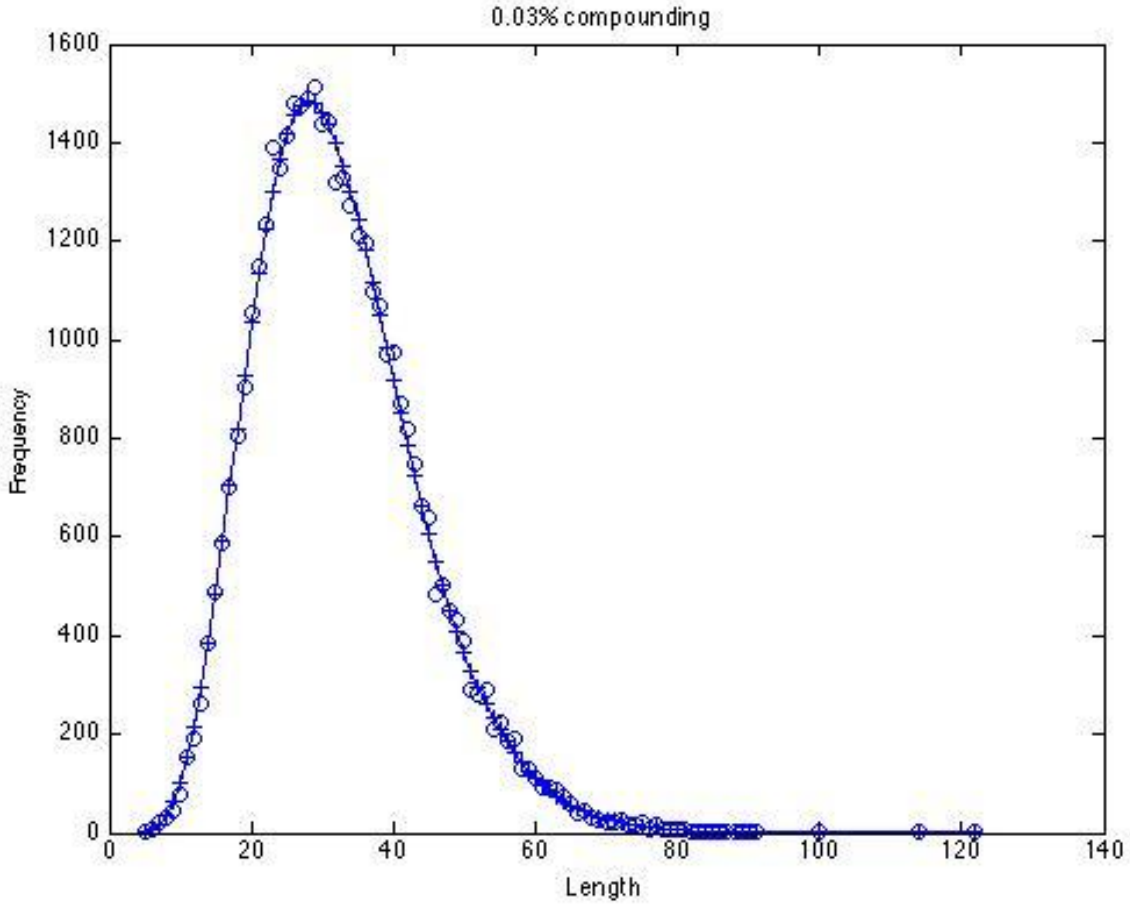


Figure 12. Simulation with 0.03% compounding

Thus, it seems that a redistribution process that includes a limited amount of compounding still results in an equilibrium fitting the Good distribution.

Word boundary shifting is but one of several length changing processes occurring in the development of language. For an overview of various lengthening processes, see Sigurd (1987). Shortening (by abbreviation) is treated in Sigurd (1979). The fact that the simulations can accommodate some compounding is encouraging, opening up for the inclusion of other such length changing processes.

5. Conclusions

Data from several genetically and typologically distinct languages indicate that the length in phonemes or letters of the words in a vocabulary is well approximated by the Good distribution. In as much as letters can be regarded as approximations of phonemes, this makes the Good distribution a candidate for a word length universal. It is suggested that this is the result of repeated lexical restructuring made in the process of oral tradition of language.

The fit is relatively insensitive to the definition and delimitation of the vocabulary units. Length can be measured in letters as well as in phonemes. The units can be restricted to base forms but also include inflected forms and list-like compounds.

So far, the data are mostly Latin alphabet orthographic word lists. It would be interesting to include more phonetic data, as well as data written in non-Latin scripts.

Computer simulations of the restructuring process result in equilibrium distributions close to the Good distribution. It remains to be proved mathematically that this is always the case. The fact that some compounding can be included in the simulations opens up for extended models which can accommodate not only word boundary shifts but also various forms of single-word shortening and lengthening known from historical linguistics.

Acknowledgements
I am indebted to Bengt Sigurd and Gabriel Altmann for valuable comments on an earlier version of this paper.

References

Alekseev, P.M. (1998). Graphemic and Syllabic Length of Words in Text and Vocabulary. *Journal of Quantitative Linguistics 5(1-2), 5-12*.

Allén, S. (1970). *Nusvensk frekvensordbok baserad på tidningstext. 1: Graford, homograf-komponenter* (Frequency Dictionary of Present-Day Swedish Based on Newspaper Material). Stockholm: Almqvist & Wiksell.

Barnes, M.R. (1980). A Nadder / An Adder: The Nasal shift. *Neophilologus 64(1), 109-112.*

Best, K.-H. (2009). *Bibliographische Übersicht zum Göttinger Projekt zur Quantitativen Linguistik.* (Stand 14.5.2009). Website: http://www.gwdg.de/~kbest/litlist.htm accessed 2009-06-12

Chatterjee, A., Chakrabarti, B.K. (2007). Kinetic exchange models for income and wealth distributions. *The European Physical Journal B 60, 135-149.*

Eeg-Olofsson, M. (2008). Why is the Good distribution so good? Towards an explanation of word length regularity. *Working Papers 53, 15-21*. (Department of Linguistics and Phonetics, Lund University).

Halácsy, P., Kornai, A. Németh, L. Rung, A., Szakadát, I., Trón, V. (2004). Creating open language resources for Hungarian. In: *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004).*

Hedelin, P. (1997). *Norstedts Svenska Uttalslexikon.* Stockholm: Norstedts Ordbok.

Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika 45(1/2), 222-228.*

Hungarian webcorpus. Website: http://mokk.bme.hu/resources/webcorpus/ accessed 2009-04-08

Johnson, N.L., Kemp, A.W., Kotz, S. (2005). *Univariate Discrete Distributions.* Third Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D. (2006). Web-based frequency dictionaries for medium density languages. In: A. Kilgarriff, M. Baroni (eds.), *Proceedings of the 2nd International Workshop on Web as Corpus:* 1-9.

Kučera, H., Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Kulasekera, K.B., Tonkyn, D.W. (1992). A new discrete distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation 21(2), 499-518.*

Leipzig Corpora Collection. Website: http://corpora.uni-leipzig.de/download.html accessed 2007-03-11

Lönngren, L. (ed.) (1993). *Chastotnyj slovar' sovremennogo russkogo jazyka* (A Frequency Dictionary of Modern Russian. With a Summary in English). *Acta Universitatis*

Upsaliensis, Studia Slavica Upsaliensia 32.

Lupsa, D.A., Lupsa, R. (2005). The law of word length in a vocabulary. *Studia univ. Babeş-Bolyai, Informatica L: 2, 69-80.*

MRC Psycholinguistic Database: Machine Usable Dictionary. Website: http://www.psych.rl.ac.uk/MRC_Psych_Db.html. accessed 2009-04-21

Ottevanger, W. *MATLAB implementation of the polylog function*. Website: http://www.mathworks.com/matlabcentral/fileexchange/23060?controller=file_infos&download=true - accessed 2009-02-23

Patriarca, M., Anirban Chakraborti, A., Kaski, K. (2004). Gibbs versus non-Gibbs distributions in money dynamics. *Physica A* 340, 334-339.

Regress+. Website: http://www.causascientia.org/software/Regress_plus.html accessed 2007-03-11.

Sigurd, B. (1979). Förkortningarna och det moderna samhället. *Språkvård 2, 3-8.*

Sigurd, B. (1987). Språkliga förlängningskonster. *Språkvård 3, 18-23*.

Sigurd, B., Eeg-Olofsson, M., Weijer, J.v.d., Strömqvist, S. (forthcoming). The most frequent words: meanings, lengths, access times. To appear in *Studia Linguistica*.

Sigurd, B., Eeg-Olofsson, M., Weijer, J.v.d. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica 58(1), 37-52.*

Uppsala Russian Corpus. Website: http://www.slaviska.uu.se/korpus.htm accessed 2009-02-17

Weisstein, E.W. Polylogarithm. From: MathWorld - A Wolfram Web Resource. Website: http://mathworld.wolfram.com/Polylogarithm.html accessed 2009-03-21

Wessén, E. (1979). *De nordiska språken*. Stockholm: Almqvist & Wiksell Förlag AB.

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: Walter de Gruyter.

Appendix 1

Perl code implementing the redistribution process

```
sub redist { # Simulation of redistribution process
# All lengths are natural numbers
# Parameters: Number of steps, propensity factor, reference to data array
# Updates the data array, which must have been initialized previously

        my ($steps, $save, $lex) = @_;

        my $vocab = scalar @$lex; # Number of units

        foreach (1..$steps) {
                my $lex1 = int(rand($vocab)); # Index of random unit 1
                my $lex2 = int(rand($vocab)); # Index of random unit 2
                while ($lex2 == $lex1) { # The units must be distinct
                        $lex2 = int(rand($vocab));
                }
                my $old1 = $$lex[$lex1]; my $old2 = $$lex[$lex2]; # Initial lengths of units
                my $total = $old1 + $old2; # Total length - does not change
                my $save1 = int(0.5 + $save*$old1); # To be saved by unit 1, rounded to
```

```
nearest integer
                my $save2 = int(0.5 + $save*$old2); # To be saved by unit 2, rounded to
nearest integer
                my $common = $total - $save1 - $save2; # To redistribute randomly
                my $new1 = int(0.5 + rand(1)*$common + $save1); # New length for unit 1
                my $new2 = $total - $new1; # New length for unit 2
                $$lex[$lex1] = $new1; $$lex[$lex2] = $new2; # Updating data array with new
values
        }
}
```

Glottometrics 19, 2009, 70-79

Diversification of postpositions in Japanese

Haruko Sanada, Tokyo
Gabriel Altmann, Lüdenscheid

Abstract. In this study we show the place of the diversification of Japanese postpositions in the hierarchy of different language phenomena. The Popescu indicator is computed from three different texts.

Keywords: *Japanese, diversification, rank-frequency sequence, Popescu indicator*

1. Introduction

In Japanese, many grammatical functions are represented by postpositions, being an analytic means, sharing this role with otherwise agglutinative constructions. Some Ugro-Finnish languages have instead affixes. The Japanese postpositions have been classified by the National Language Research Institute (cf. NLRI 1951) as follows:

case postposition	(kaku joshi)
linking postposition	(kei joshi)
ending postposition	(shu joshi)
attributive postposition	(juntai joshi)
conjunctive postposition	(setsuzoku joshi)
adverbial postposition	(fuku joshi)
interjectional postposition	(kanto joshi)
listing postposition	(heiritsu joshi)

Since inflections, affixes or prepositions of other languages have frequently more than one function or meaning (cf. e.g. Nemcová 2007), it can be expected that in Japanese, too, at least some postpositions are not mono-functional or mono-semantic. For example, the Japanese *ka* can be adverbial, listing and ending. But even in one special function these auxiliary elements can be polysemic. This can be seen especially if one tries to translate sentences from one language into another. For example, the English preposition *in* can be translated into German *in, an, auf, innerhalb, unter, während, binnen, bei, zu, nach, gemäß, aus, von, innen, drinnen* etc. It can be translated into Japanese as *no naka ni, no naka de, no naka no, no aida, no ato ni, dewa, no, ni, shite, de, no tame ni,* etc. Thus considering postpositions we have two dimensions: the grammatical function – usually classified in more or less traditional way with a necessary extent of fuzziness – and the polysemy of each of these functions. A very thorough, extensive work would be necessary to capture this two-dimensional structure quantitatively because no tagged corpora can contain everything. From the quantitative point of view we would obtain a two-dimensional frequency distribution. Would we add also the allomorphs (different forms) of individual postpositions, we would obtain a three-dimensional structure, but our problem here is different.

According to a hypothesis of quantitative linguistics, if a class of entities is correctly set up, then the rank-frequency distribution of the entities abides by an adequate probability distribution or regression function (cf. Strauss, Fan, Altmann 2008:94). In addition, the

empirical distribution displays some measurable characteristic features which possibly give a hint at the place of the given phenomenon in the classification of linguistic phenomena. However, in counting frequencies we are confronted with the fact that a corpus is a mixture of texts. Its immense size does not lead in any case to the stabilization of frequencies of all phenomena and at the same time testing with classical statistics is made almost impossible. This is why we shall use here different texts to measure the frequencies of postpositions. We suppose that not all of them occur in each of our texts but that their rank-frequency distributions have the same properties.

In the Japanese grammar the following postpositions are reported (NLRI 1951) including colloquial forms marked with "*" in the list:

ba	*made*	*tara**/*ttara**
bakari	*mo*	*tari* (*dari* after a nasal syllable)
bakoso	*mono*	*tsutsu*
dake	*monoka/monka**	*tte** (contracted from *to* and *wa*)
*datte**	*monode/monde**	*te* (*de* after a nasal syllable)
*dano**	*mononara*	*teba**/*tteba**
de	*monono*	*temo* (*demo* after a nasal syllable)
demo	*monoo*	*to*
dokoroka	*na**	*toka**
domo	*nagara*	*tokoroga*
e (case, "へ")	*nado/nazo**/*nanzo**/*nanka**	*tokorode*
e (ending, "え")	*nari*	*tote*
hodo	*nante**	*tomo*
*i**	*ni*	*wa* (linking, "は")
ga	*ne**/*nee**	*wa** (ending, "わ")
ka	*no/n**	*ya*
*kashira**	*node*	*yara*
kara	*noni*	*yo**
ke	*nomi*	*yori*
*keredomo/keredo**/*kedo**/*kedom o**	*o*	*ze**
kiri	*sa**	*zo**
koso	*sae*	*zutsu*
koto	*shi*	
kototote	*shika*	
*kuseni**	*shimo*	
kurai/gurai	*sura*	
	*tatte** (*datte** after a nasal syllable)	

2. The rank-frequency sequence of postpositions

Analyzing three texts, namely *Jinseiron Note* (essay, Miki 1941, 1995), *Kusa no hana* (novel, Fukunaga 1954, 1995) and *Shinbashi Karasumori Guchi Seishunhen* (novel, Shiina 1987, 1995) we obtain the frequency results as shown in Table 1.

Table 1
Frequencies of postpositions in individual texts

Jinseiron Note			Kusa no hana		Shinbashi Karasumori Guchi	
Order	Postpos.	Freq.	Postposition	Freq.	Postposition	Freq.
1	no	2614	no	4279	no	4399
2	wa	2076	wa	3534	te/de	2747
3	ni	1832	te/de	3270	ni	2016
4	te/de	1493	ni	3124	to/tto*	1945
5	to	1239	o	2887	o	1938
6	ga	1127	ga	2366	ga	1901
7	o	929	to/tto*	1975	wa	1892
8	mo	477	mo	1320	de	988
9	ka	347	de	830	mo	877
10	kara	266	no	780	ka	662
11	de	190	ka	632	kara	458
12	ba	174	kara	616	ya	302
13	yori	69	n* (<-no)	452	yo*	298
14	nomi	44	yo*	344	to	282
15	ya	42	ne*/nee*	271	n* (<-no)	237
16	hodo	41	to	238	na*/naa*	209
17	sae	36	ba	186	ne*/nee*	185
18	nagara	30	made	179	nagara	159
19	made	26	tte*	175	hodo	107
20	e	22	e (case)	171	tte*	101
21	dake	22	dake	132	keredo/kedo*	96
22	tomo	15	nagara	117	nado	93
23	koso	14	ya	114	dake	90
24	shika	11	nanka*	97	made	89
25	tsutsu	8	na*/naa*	79	sa*/saa*	61
26	nado	8	hodo	69	ba	54
27	shi	7	i*	69	kurai/gurai	54
28	na	7	bakari/bakkashi*/ bakkari*	67	yori	53
29	bakari	5	nado/nazo*/nanzo*	65	tari/dari	44
30	tari	4	shi	64	e (case)	41
31	shimo	4	tari/dari	61	shika	37
32	sura	3	keredo/kedo*	57	shi	36
33	zo*	2	yori	55	bakari	30

34	ne	1	sa*/saa*	53		no	25
35	n	1	sae	41		zo*	24
36			nante*	39		ja*/jaa* (<-dewa)	22
37			kashira	31		nomi	16
38			kiri/kkiri*	31		ke*/kke*	14
39			shika	27		zutsu	14
40			cha* (<-tewa)	24		ze*	13
41			tatte*/ttatte*	23		nante	12
42			zutsu	23		wa	11
43			kurai/gurai	22		nanka*	10
44			ja* (<-dewa)	21		i*	9
45			ze*	20		mono/mon*	6
46			tsutsu	16		tatte*/ttatte*	6
47			koso	15		tsutsu	6
48			mono	13		cha* (<-tewa)	5
49			tomo	12		shimo	4
50			nomi	10		kashira	3
51			simo	7		koso	3
52			zo*	7		sae	3
53			ke*/kke*	5		kiri/kkiri*	2
54			yara	5		sura	2
55			nari	2		yara	2
56			nite	2		domo	1
57			ttara*	1		nari	1
58						tomo	1

In Table 1 we have a rank-frequency distribution of a selected set of words. As can be seen at the beginning, they form eight sets mentioned above some of which display intersections. Even if we eliminate the intersections and retain only one of the identical elements (having different functions) it is not possible to fit any of the known distributions to this data set. It is caused by two circumstances (1) the sets do not cover the whole text, only a part of it, (2) they form strata of sets, each having its own distribution; if we considered each set separately, some of the "Zipfian" distributions could capture the data. However, the sets have a weak intersection, thus we can consider them as complementing one another in different functions. In that case it is possible to overlay the individual strata, re-rank the units and consider it a sequence of numbers. According to a proposal of I.-I. Popescu, Altmann, Köhler (2009) such a superposition can be captured by the function

$$(1) \qquad y = 1 + ae^{-bx} + ce^{-dx} + \dots$$

and if there is a certain harmony (complementation) between the strata (as is e.g. between vowels and consonants, autosemantics and synsemantics etc.), then one of the components of formula (1) is sufficient to capture the rank-frequency sequence of data. The above function has the advantage to signalize how complex the superposition is, in that it yields equal

exponents if the mixture is overestimated. Computing (1) for the data in Table 1 we obtain the theoretical results in Table 2. As can be seen, one component is sufficient to yield good fitting. The results are presented graphically in Figures 1, 2 and 3.

Table 2
Fitting function (1) to postposition data.

Rank	Jinseiron Note		Kusa no hana		Shinbashi Karasumori Guchi	
r	f_r	$\hat{y}$	f_r	$\hat{y}$	f_r	$\hat{y}$
1	2614	2702.95	4279	4531.84	4399	3844.26
2	2076	2172.93	3534	3834.68	2747	3182.28
3	1832	1746.89	3270	3244.78	2016	2634.31
4	1493	1404.41	3124	2745.66	1945	2180.74
5	1239	1129.12	2887	2323.34	1938	1805.29
6	1127	907.83	2366	1966.00	1901	1494.51
7	929	729.94	1975	1663.64	1892	1237.26
8	477	586.95	1320	1407.81	988	1024.32
9	347	472.01	830	1191.34	877	848.06
10	266	379.62	780	1008.18	662	702.15
11	190	305.35	632	853.20	458	581.38
12	174	245.65	616	722.08	302	481.41
13	69	197.66	452	611.12	298	398.67
14	44	159.08	344	517.24	282	330.17
15	42	128.07	271	437.81	237	273.47
16	41	103.14	238	370.60	209	226.54
17	36	83.11	186	313.73	185	187.69
18	30	67.00	179	265.61	159	155.53
19	26	54.05	175	224.89	107	128.92
20	22	43.65	171	190.44	101	106.88
21	22	35.28	132	161.29	96	88.65
22	15	28.56	117	136.63	93	73.55
23	14	23.15	114	115.76	90	61.05
24	11	18.81	97	98.10	89	50.71
25	8	15.31	79	83.16	61	42.15
26	8	12.51	69	70.52	54	35.06
27	7	10.25	69	59.82	54	29.19
28	7	8.43	67	50.77	53	24.34
29	5	6.98	65	43.11	44	20.32
30	4	5.80	64	36.63	41	16.99
31	4	4.86	61	31.15	37	14.24
32	3	4.10	57	26.51	36	11.96

33	2	3.50	55	22.58	30	10.07
34	1	3.01	53	19.26	25	8.51
35	1	2.61	41	16.45	24	7.21
36			39	14.08	22	6.14
37			31	12.06	16	5.26
38			31	10.36	14	4.52
39			27	8.92	14	3.92
40			24	7.70	13	3.41
41			23	6.67	12	3.00
42			23	5.80	11	2.65
43			22	5.06	10	2.37
44			21	4.44	9	2.13
45			20	3.91	6	1.94
46			16	3.46	6	1.78
47			15	3.08	6	1.64
48			13	2.76	5	1.53
49			12	2.49	4	1.44
50			10	2.26	3	1.36
51			7	2.07	3	1.30
52			7	1.90	3	1.25
53			5	1.76	2	1.21
54			5	1.65	2	1.17
55			2	1.55	2	1.14
56			2	1.46	1	1.12
57			1	1.39	1	1.10
58					1	1.08
	a = 3361.3072 b = 0.2184 R^2 = 0.9851		a = 5354.7916 b = 0.1671 R^2 = 0.9785		a = 4642.9967 b = 0.1890 R^2= 0.9592	

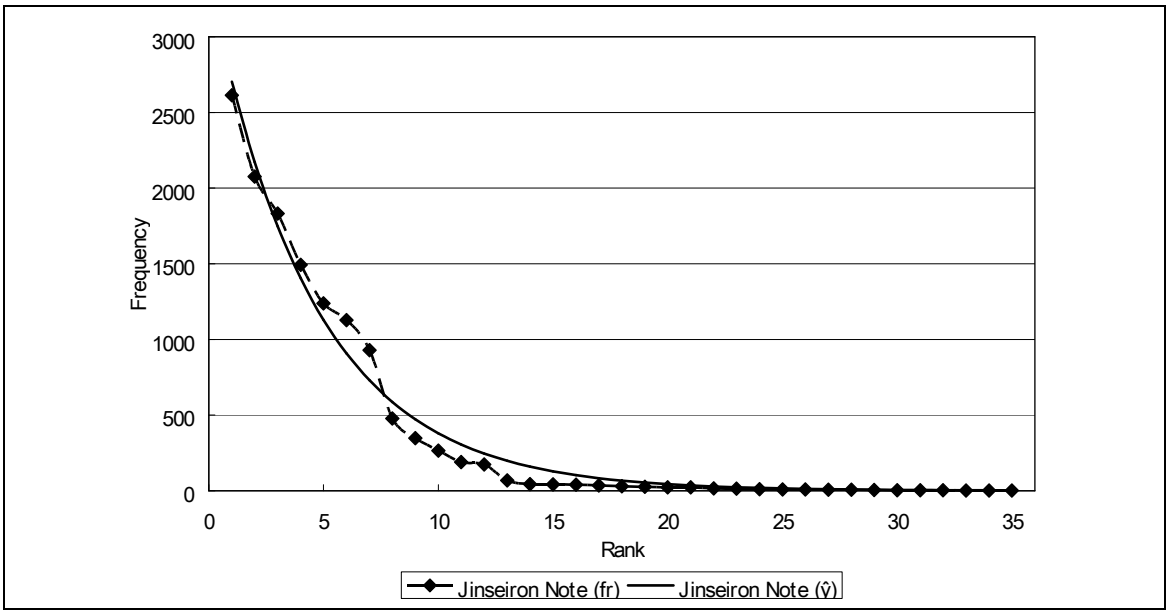


Figure 1. Fitting (1) to Jinseiron Note

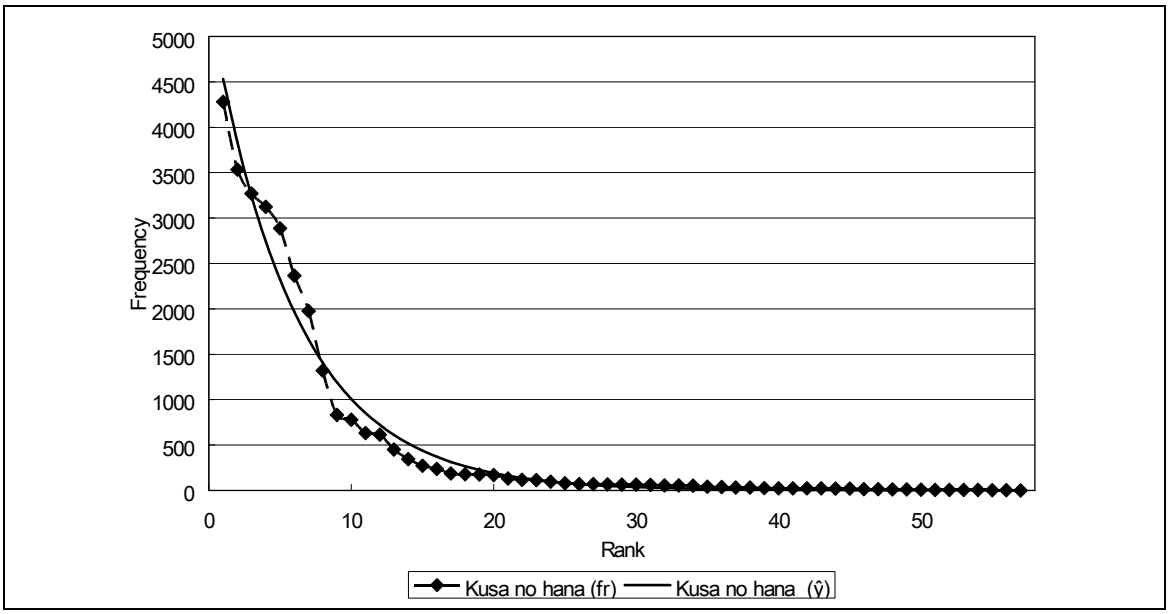


Figure 2. Fitting (1) to Kusa no hana

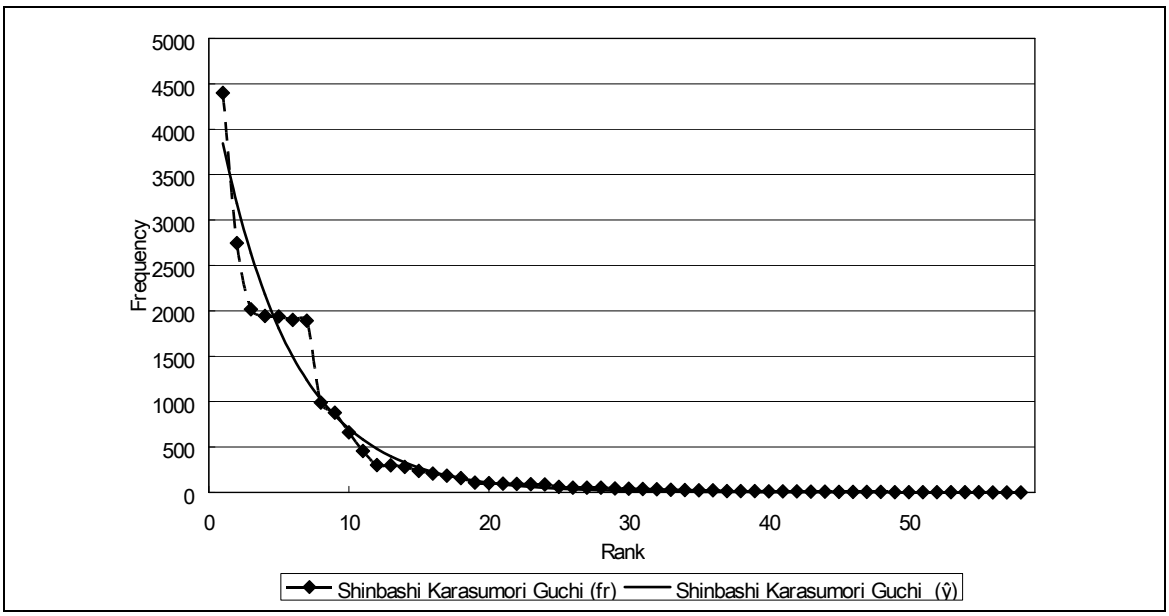


Figure 3. Fitting (1) to Shinbashi

3. Diversification

Once a language discovers an efficient phonemic, grammatical, semantic etc. means for effective communication, the given means begins to diversify (cf. Köhler 1991). But the diversification is not quite haphazard, it develops very regularly, but differently for different language phenomena. The direct causes of diversification are not known and it may develop in different dimensions, e.g. an affix can get different forms, it can attain different meanings, it can be combined with different word classes, it can obtain different grammatical functions, etc.

In order to capture this regularity, Popescu and Altmann (2008) proposed a coefficient which is very stable for a given phenomenon in different languages. The coefficient *c* can be computed as follows:

$$c = \frac{R + f_{\max} - f_{\min} + 1 - L}{h}. \tag{2}$$

Here, R is the number of different postpositions in a text, as a matter of fact, the inventory which is identical with the maximum rank r_{max}. Further f_{max} is the frequency at the first rank, i.e. f_1 ; f_{min} is the frequency at the last rank, if. $f_{min} = f_R$; L is the length of the arc between the empirical values computed as the sum of Euclidean distances between individual f_r-s i.e.

$$L = \sum_{r=1}^{R-1} [(f_r - f_{r+1})^2 + 1]^{1/2} \tag{3}$$

and h is the fixed point which can be computed in different ways, we adhere here to the simple formula

$$(4)\quad h = \begin{cases} r & \text{if there is an } r = f_r \\ \dfrac{f_a r_b - f_b r_a}{r_b - r_a + f_a - f_b} & \text{if there is no } r = f_r \end{cases}$$

in which we either consider h such an r which is equal to f_r – unfortunately not present in our data - or we take the values $f_a > r_a$ and $f_b < r_b + 1$. Usually $r_b = r_a + 1$. For example, in Jinseiron Note $r_a = 21$, $f_a = 22$, $r_b = 22$, $f_a = 15$, from which $h = 21.125$. In many cases one simply takes the rounded value. The computation of L is simple. For our three texts we obtain the results in Table 3.

Table 3
Indicator c for three Japanese texts

Text	R	f_1	f_R	L	h	c
Jinseiron Note	35	2614	1	2622.2833	21.21	1.26
Kusa no hana	57	4270	1	4291.0532	36.33	1.24
Shinbashi	58	4399	1	4416.0280	32.57	1.26

As can be seen, the indicator c is very stable in Japanese. It is to be mentioned that U. Roos (1991) analyzed the semantic diversication of the postposition *ni*, i.e. the diversification of one unit in semantic dimension, and Popescu, Altmann (2008) obtained for it $c = 1.35$. The preliminary result for prepositions, conjunction and postpositions in several languages is given by Popescu and Altmann as $\overline{c} = 1.24$ and the 95% confidence interval for individual values is <1.03, 1.46>. As can easily be seen, the values of Japanese postpositions taken as a whole have very similar values which in any case lie in the prescribed interval. The semantic diversification is slightly greater and should be analyzed for each postposition separately. As has been observed, semantic diversifications have usually greater c-values then formal ones. For meaning diversification of affixes Popescu and Altmann report $\overline{c} = 1.39$ and for semantic diversification of words $\overline{c} = 1.47$ (cf. e.g. Fan, Altmann 2008). The coefficient c can be considered as an indicator of the adequateness of forming a class of entities (cf. also Popescu, Mačutek, Altmann 2008; Fan, Popescu, Altmann 2008).

Another indicator proposed by Popescu, Mačutek and Altmann (2008a) is p computed as

$$(5)\quad p = \frac{L_{\max} - L}{h - 1}$$

which is easy to compute because

$$(6)\quad L_{max} = (R - 1) + f(1) - f(R),$$

hence it is merely a slight modification of c. For our three data we obtain

	p
Jinseiron Note	1.23
Kusa no hana	1.22
Shibashi	1.23

Popescu, Mačutek and Altmann (2008) give as an interlingual mean $\overline{p}$ = 1.131 and the 95% confidence interval for individual *p*-values as <0.88, 1.39>. As can be seen, the Japanese postpositions lie in the given interval, even if all of them are slightly greater than the general mean. Meaning diversification of postposition *ni* is estimated from Roos (1961) to $p = 1.24$.

One of these two indicators is sufficient to show that phenomena at different language levels diversify but all of them have their proper way which can be expressed quantitatively. The indicators show the pace of the development.

Newer analyses using the German language of SMS on many different phenomena were performed by Laufer and Nemcová (2009).

References

Fan, F., Altmann, G. (2008). On meaning diversification in English *Glottometrics 17, 2008, 66-78.*

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

Fukunaga, T. (1954, 1995). *Kusa no hana* (Flowers of herbs). Tokyo: Shinchôsha. (included in Shinchôsha 1995).

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (ed), *Diversification Processes in Language: Grammar: 47-55*. Hagen: Rottmann.

Laufer, J., **Nemcová, E.** (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25.*

Miki, K. (1941, 1995). *Jinseiron Note* (Essay on the life). Tokyo: Kawade shobô. (included in Shinchôsha 1995).

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.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 94-108.*

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf´s law – another view. *Quality and Quantity (submitted).*

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

Popescu, I.-I., Mačutek, Altmann, G. (2008a). *Aspects of word frequencies*. Lüdenscheid: RAM.

Popescu et al. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Roos, U. (1991). Diversifikation der japanischen Postposition "-ni". In: Rothe, U. (ed), *Diversification Processes in Language: Grammar: 75-82*. Hagen: Rottmann

Shiina, M. (1987, 1995) *Shinbashi Karasumori Guchi Seishunhen* (Karasumori exit of Shinbashi station: Part of youthful days). Tokyo: Shinchôsha. (included in Shinchôsha 1995).

Shinchôsha (1995). Shinchô bunko no 100 satsu, CD-ROM ban (Selected 100 paperbacks of Shinchôsha, CD-ROM edition).

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in quantitative linguistics 1*. Lüdenscheid: RAM.

The National Language Research Institute (1951). *Gendaigo no joshi jodoshi: yoho to jitsurei.* (Bound forms in modern Japanese - Uses and examples). Tokyo: Shuei shuppan.

Glottometrics 18, 2009, 80-84

Zur Entwicklung der Entlehnungen aus dem Japanischen ins Deutsche

Karl-Heinz Best

Abstract. This study presents a further support of the logistic law, known in linguistics as Piotrowski law, using data which can be gathered from *Kleines Lexikon deutscher Wörter japanischer Herkunft* (2008).

Keywords: German, Japanese, borrowings

1. Entlehnungen ins Deutsche

In diesem Beitrag geht es um die Entlehnungen aus dem Japanischen ins Deutsche. 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

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

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 Entlehnungen aus dem Japanischen werden alle die Wörter betrachtet, die entweder direkt aus dem Japanischen als Herkunftssprache stammen oder über andere Sprachen als Vermittlersprachen ins Deutsche gekommen sind. Zwischen Fremd- und Lehnwörtern wird nicht unterschieden; einzelne Lehnübersetzungen wurden berücksichtigt.

2. Die Datengrundlage

Die Gelegenheit zu dieser Untersuchung eröffnet das Buch von Haschke & Thomas (2008), das einen guten Überblick über die Entlehnung japanischer Wörter ins Deutsche bietet. Die Autoren geben an, dass in neuen Wörterbüchern des Deutschen ca. 500 Wörter japanischer Herkunft zu finden seien und beziffern ihren Anteil am deutschen Wortschatz auf „0.3 Prozent der Einträge im Duden-Fremdwörterbuch" (Haschke & Thomas 2008: 6); dies stimmt grob mit den Angaben überein, die Best (2001, 2005) und Körner (2004) vorgelegt haben, wo die japanischen Entlehnungen auf 1 bzw. 6% beziffert wurden.

Die Untersuchung von Haschke & Thomas ist trotz dieser Übereinstimmung von Interesse, da der erfasste Wortschatz ein Vielfaches dessen ist, was in den früheren Untersuchungen berücksichtigt werden konnte. Hinzu kommt, dass die neue Untersuchung für jedes Wort angibt, in welchen Wörterbüchern es auftaucht. Man muss also nur für jedes einzelne Wort herausfinden, wann es zu ersten Mal belegt ist, also: welches das älteste Wörterbuch ist, in dem ein bestimmtes Wort japanischer Herkunft genannt ist. Da die älteste namentlich ge-

nannte Quelle das Fremdwörterbuch von Heyse (1901) ist, mussten einige Wörter neu datiert werden.

Nach Haschke und Thomas (2008: 9) sind bei Heyse als Wörter japanischer Herkunft aufgeführt: „Aucuba, Bonze, Geisha, Japan, Mikado, Nipon und Soja." Nach Auskunft etymologischer Wörterbücher ist Bonze eine Entlehnung des 16. Jhds. (Kluge 2002), Soja stammt aus dem 18. Jhd. (Kluge 2002), Mikado aus dem 19. Jhd. (*Duden. Herkunftswörterbuch* 2001). Hinzu kommen Harakiri und Kimono, die laut *Duden. Herkunftswörterbuch* (2001) bereits im 19. Jahrhundert ins Deutsche gelangten, sowie Gingko, das spätestens mit Goethes Gedicht *Gingo biloba* von 1815 als bekannt angenommen werden kann. Das Wort „Japan" wurde allerdings nicht als Entlehnung aus dem Japanischen betrachtet; dieses Wort ist ein Exonym, d.h. eine nicht einheimische Bezeichnung für das Land, während im Japanischen selbst stattdessen „Nihon", „Nipon" bzw. „Nippon" verwendet wird. Die Bezeichnung „Japan" fehlt nach Haschke & Thomas (2008: 77) außerdem in japanischen Wörterbüchern; sie wurde in der Statistik der Entlehnungen daher nicht berücksichtigt, ebenso wie die Wörter, die von „Japan" abgeleitet wurden oder an denen dieses Wort als Konstituente eines Kompositums beteiligt ist. Entsprechend diesen Hinweisen wurde der Wortschatz, der Heyse (1901) zugewiesen ist, gekürzt.

Bei der Aufnahme der Wörter wurde außerdem darauf geachtet, dass dann, wenn ein Wort im Deutschen in verschiedenen Schreibweisen auftaucht, nur der Erstbeleg berücksichtigt wurde. Ein solcher Fall ist etwa das Wort für Japans berühmten Berg, der bei Haschke & Thomas als „Fudschi", „Fudschijama", „Fudschi-no-yama" und „Fuji(-san)" erscheint und nur einmal gewertet wurde. Im Falle von „Daimio" , „Daimjo" und „Daimyo" wurden „Daimio"/"Daimyo" als ein Beleg des Zeitraums 1928-34 gewertet; das wesentlich später belegte „Daimyo" nicht.

Da für die Zeit bis zum beginnenden 20. Jahrhunderts nur ganz wenige Entlehnungen aus dem Japanischen belegt sind, wurde als erster Zeitraum das 16. – 19. Jahrhundert mit insgesamt 6 Wörtern angesetzt. Die weiteren Datierungen richten sich nach den Angaben von Haschke & Thomas, was nicht ohne eine gewisse Willkür möglich ist. Die Zeitangabe 1935 bezieht sich auf die Wörter, die erstmals im 20bändigen Brockhaus von 1928-34 und in Kimura verzeichnet sind. Haschke und Thomas beziehen sich zwar auf Kimura (1961 und später); dieses Wörterbuch erscheint jedoch ab ca. 1920 und hat bereits in den 1930er Jahren genau den gleichen Umfang. Das Jahr 1965 steht für das *Fremdwörterbuch* (Leipzig 1965), 1980 für Schinzinger (1980 und später), 1985 für das *Große Fremdwörterbuch* (Leipzig 1985). Beide Leipziger Ausgaben sind nicht Erstauflagen, so dass auch ein etwas früherer Ansatz gerechtfertigt wäre. Die Angabe „um 2000" schließlich versammelt *Duden, Fremdwörterbuch* (2001), *Duden. Das große Fremdwörterbuch* (2000), *Duden. Deutsches Universalwörterbuch* (2001) und Brockhaus aktuell (nicht genau spezifiziert; im Literaturverzeichnis wird auf „Der große Brockhaus (verschiedene Auflagen)" und „Der Brockhaus multimedial 2004-2006" hingewiesen). „Brockhaus" ist damit genau genommen gar nicht datiert – insofern und aufgrund der Zusammenfassung für das Jahr 2000 stecken in den Daten Verzerrungen. Der Trend der Übernahme von Entlehnungen dürfte aber dennoch einigermaßen korrekt erfasst sein.

Hinzu kommen zwei Wörter, die mangels Quellenangabe nicht datiert werden konnten: Oribe-yaki und Wapuro.

3. Der Trend der Übernahme japanischer Wörter ins Deutsche

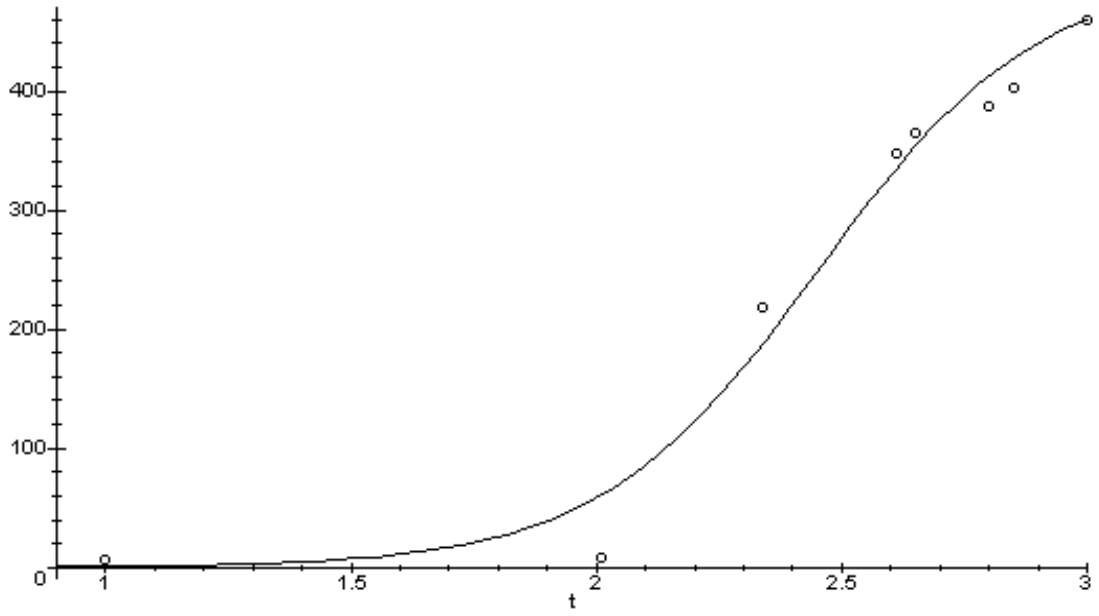
In der folgenden Tabelle 1 ist nun nach Maßgabe der Bearbeitungshinweise zusammengestellt, wann wie viele Wörter japanischer Herkunft ins Deutsche gelangt sind. Dabei ist zu

berücksichtigen, dass die meisten dieser Wörter direkt aus dem Japanischen und nur wenige über dem Umweg über andere Sprachen ins Deutsche übernommen wurden (Haschke & Thomas: 6). Die für jeden Zeitpunkt/Zeitraum ermittelten neuen Wörter („beobachtet“) wurden aufsummiert („kumuliert“) und an diese kumulierten Daten wurde das Piotrowski-Gesetz in der Form für den unvollständigen Sprachwandel (Formel (1)) angepasst; die so ermittelten Werte für die einzelnen Zeiten sind unter „berechnet“ aufgeführt:

Tabelle 1
Japanische Entlehnungen im Deutschen

Zeit	*t*	beobachtet	kumuliert	berechnet
16. - 19. Jahrhundert	1	6	6	2.33
1901	2.01	3	9	100.15
1935	2.35	339	348	244.39
1965	2.65	17	365	378.57
1980	2.80	22	387	424.58
1985	2.85	15	402	436.35
um 2000	3	80	460	462.63
a = 10977.3896 *b* = 3.9399 *c* = 500 *D* = 0.90				

Legende: *a*, *b* und *c* sind die Parameter des Modells; *c* gibt an, auf welchen Zielwert der Prozess hinausläuft. Hier ist *c* in Übereinstimmung mit Haschke & Thomas (2008: 8) mit *c* = 500 angesetzt worden. Das Ergebnis der Anpassung ist mit *D* = 0.90 sehr gut; die erwähnten – vermutlich nicht allzu großen – Verzerrungen in den Daten haben jedenfalls nicht dazu geführt, dass man Modell (1) für den Prozess der Übernahme japanischer Wörter ins Deutsche verwerfen müsste. Die folgende Graphik verdeutlicht noch einmal das gute Ergebnis:



Graphik zu Tabelle 1: Japanische Entlehnungen im Deutschen (*t* = 1 steht für das 19. Jahrhundert, *t* = 2.01 für das Jahr 1901, etc.)

4. Zusammenfassung

Obwohl mit gewissen zeitlichen Verzerrungen in den Daten gerechnet werden muss, kann durch Anpassung von Formel (1) ein deutlicher Trend für die Übernahme von japanischen Wörtern ins Deutsche dargestellt werden. Je nach der Art der Verzerrung könnte die berechnete Kurve ein wenig flacher oder auch steiler verlaufen. Am Gesamtbild würde sich kaum etwas ändern.

Das Japanische ist in den weitaus meisten Fällen die Sprache, aus der die übernommenen Wörter herstammen (Herkunftssprache). Nur in wenigen Fällen sind japanische Wörter über andere Sprachen (Vermittlersprachen) ins Deutsche gelangt; hierzu gehören nach Haschke & Thomas (2008: 8) die Wörter Bonze, Soja und Tycoon. Im Falle von „Nippon", dem ein chinesisches Wort zugrunde liegt, wird deutlich, dass Japanisch auch einmal als Vermittlersprache fungieren kann.

Als weiteres wesentliches Ergebnis kann festgestellt werden, dass das Piotrowski-Gesetz offenbar auch anhand der japanischen Entlehnungen ins Deutsche gestützt werden kann.

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.*

Duden. *Deutsches Universalwörterbuch.* 4., neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2001.

Duden. *Fremdwörterbuch.* 7., neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2001.

Duden. *Das große Fremdwörterbuch. Herkunft und Bedeutung der Fremdwörter*. 2., neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2000.

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

Haschke, Barbara, & Thomas, Gothild (2008*). Kleines Lexikon deutscher Wörter japanischer Herkunft von Aikido bis Zen.* München: Beck.

Heyse, Johann Christian August (1901). *Joh. Christ. Aug. Heyses allgemeines verdeutschendes und erklärendes Fremdwörterbuch mit Bezeichnung der Aussprache und Betonung der Wörter nebst genauer Angabe ihrer Abstammung und Bildung: unter Berücksichtigung der amtlichen Erlasse über Verdeutschung der Fremdwörter und der neuen einheitlichen Rechtschreibung.* 19. Auflage, rev. und verm. v. Ed. Loewenthal. Berlin: Cronbach.

Fremdwörterbuch. 9., verb. Aufl. Leipzig: VEB Bibliographisches Institut 1965. (Erstauflage 1954; die 6. Auflage ist etwas erweitert.)

Großes Fremdwörterbuch. 6., durchges. Aufl. Leipzig: VEB Bibliographisches Institut 1985. (Erstauflage 1977; die 6. Auflage ist geringfügig erweitert.)

Kimura, Kinji (1936 und spätere Auflagen). *Großes Japanisch – Deutsches Wörterbuch.* Tokio: Hakubunkan.

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.*

Schinzinger, Robert (Hrsg.) (1980 und spätere Auflagen). *Wörterbuch der deutschen und japanischen Sprache.* Tokio: Sansyusha.

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 19, 2009, 85-98

Word form and lemma syntactic dependency networks in Czech: a comparative study

Radek Čech[1], *Ostrava*
Ján Mačutek[2], *Graz*

Abstract. We compare several parameters of word form and lemma syntactic dependency networks in Czech. Models for degree distributions are suggested.

Keywords: syntactic dependency network, word form, lemma, degree distribution.

1. Introduction

A complex network analysis (e.g., Caldarelli 2007) has been used for several studies of human language in recent years, and now it is well known that linguistic networks display some statistical properties of complex networks (Mehler 2007). These properties have been detected at different linguistic levels – there are analyses of *semantic* networks (Sigman, Cecchi 2002; Motter et al. 2002; Holanda et al. 2004; Steyvers, Tenenbaum 2005), *syntactic* networks (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005), *co-occurrence* networks (Ferrer i Cancho, Solé 2001, Dorogostev, Mendes 2001), and *syllabic* networks (Soares, Corso, Lucena 2005). The fact that complex network features have been found across languages as well as across linguistic levels seems to suggest a new type of human language universals (Ferrer i Cancho 2005). The network approach was also applied to language development analysis (Ninio 2006; Ke, Yao 2008).

The same statistical properties of networks have been found in non-linguistic systems – in biology, ecology, the Internet, social systems and so on (see Caldarelli 2007). So, from the perspective of network analysis, language might be under the same rules or laws as many social and biological systems.

Recent research in syntax based on the network analysis was brought in connection with some important grammatical features of languages, for example projectivity (Ferrer i Cancho, 2006a, 2008), agrammatism (Ferrer i Cancho 2005) and the relation among Zipf's law, syntax, and communication needs (Ferrer i Cancho, Riordan, Bollobás 2005; Solé 2005; Ferrer i Cancho 2006b). All these analyses are based on quantitative characteristics of language.[3]

The aim of this article is to compare properties of two syntactic dependency networks based on the same language data. The first network is created by using raw word forms, the second network by using canonical word forms – lemmas. The study reveals that both networks display some properties which are typical of complex networks: small world effect, which is given by high clustering coefficient and low average path length between nodes, and

1 Department of Czech Language, University of Ostrava, Reální 5, 70103 Ostrava, Czech Republic; e-mail: radek.cech@osu.cz
2 Institut für Slawistik, Karl-Franzens Universität, Merangasse 70, 8010 Graz, Austria; e-mail: jan.macutek@uni-graz.at or jmacutek@yahoo.com
3 To our knowledge, there is one attempt to combine traditional non-statistical approach to grammar with the network analysis (Hudson 2007), but this approach is not followed in this article.

high heterogeneity; both networks are scale free. The comparison allows to investigate (1) which network properties depend specially on the fact that one uses word forms or lemmas and (2) which factors influence prospective differences between the network based on word forms (hereinafter WFN) and the network based on lemmas (hereinafter LN).

So far, most of syntactic network analyses have worked with networks in which each node represents a word form (e.g., Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005). For instance five singular forms for seven cases of the Czech noun *kluk* (*a boy*) (nominative *kluk*; genitive & accusative *kluka*; dative & local *klukovi*; vocative *kluku*; instrumental *klukem*) are represented by five different nodes. On the other hand, if one uses lemmas, all word forms are represented by only one node – for example words such as *do, does, did, done* and *doing* are word forms of the lemma *DO* while words *kluk, kluka, klukovi, kluku, klukem* and all plural forms are word forms of the lemma *KLUK* (*BOY*).

To our knowledge, only Caldeira et al. (2006) analysed a syntactic network based on lemmas, but they used *co-occurrence* syntactic network i.e., any two words were connected if they were concomitantly in one (or more) sentence. So, the present study is the first attempt to use lemmas for syntactic *dependency* network analysis.

At first sight, the possible discrepancies between WFN and LN are caused by inflection; if there are no inflected words in a language, both networks would be equal. It opens the question about the use of the network analysis for other typological characteristics of languages. But there is no straightforward influence of inflection on network properties in the syntactic dependency networks (in the sense the more inflected words in WFN the more discrepancy between WFN and LN) because syntactic relationships could also have an important impact on the properties of LN, and consequently on the differences between WFN and LN. Hence, this article is primarily focused on the exploration of all factors that have influence on LN properties in comparison to WFN. The results show that it is possible to partly hypothesize a relationship between the typological characteristics of language and network properties (average degree, clustering coefficient).

The use of WFN is the best way for analysis of global properties of syntactic networks, primarily because it reflects syntactic dependencies of words in actual language use. The concept of lemma is artificial; nothing as lemmas actually exists in a language. However, it is obvious that using LN simplifies linguistic analysis, especially in high inflectional languages as Czech (for example the lemma of the verb *JÍST* (*EAT*) could be realized by 30 different word forms), and it is not hard to imagine why an analysis of a role of certain group of words (for instance transitive verbs) is easier and purposeful in LN than in WFN. So, if one agrees that "networks are means, not the goal" (Liu 2008) for linguists, the use of LN seems to be a reasonable way for language inquiries. But for adequate linguistic analysis of LN it is first necessary to explore the relationship between WFN and LN based on the same language data. And also prospective comparative analyses of LNs based on different languages require knowledge about factors influencing properties of LN.

It has been shown (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005) that syntactic dependency networks based on word forms display some properties that are typical of complex networks (i.e., a small world effect, high heterogeneity). So first, it is necessary to explore whether LN has the properties of complex networks as well as WFN, and then possible discrepancies between both networks have to be analysed.

This article is organized as follows. Properties of syntactic networks are presented in Section 2; the comparison of syntactic networks based on word forms and lemmas is given in Section 3; and the article is closed by Discussion.

2. Data and methodology

The data used in this study come from the Prague Dependency Treebank 2.0 (hereinafter PDT) (Hajič et. al. 2006). The PDT is a Czech corpus which contains a large amount of texts with interlinked morphological, syntactic, and semantic annotation. For the present purposes we used data annotated on the analytical layer, which contains 87,913 sentences and 1,503,739 word tokens. Thanks to the PDT lemmatization it has been possible to create WFN as well as LN. The PDT consists of articles from newspapers and journals.

In constructing the networks, we follow the method developed by Ferrer i Cancho et al. (2004). This approach is based on dependency grammar formalism which defines the structure of a sentence as a set of linked lexical nodes. The links represent binary relations of dependency between nodes. The direction of the links, going from the head to its modifier, (1) reflects types of syntactic relations which determine the morphological form of the sub-ordinate word (agreement and regimen) and (2) reflects the valency properties (see Figure 1)[4].

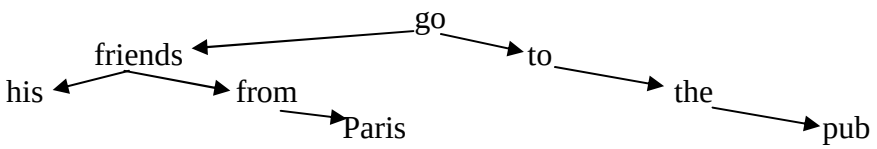


Figure 1. Direction of dependencies

The syntactic dependency network contains all words in the corpus. Two words are linked if there is a dependency relationship between them in the corpus.[5] Thus, a global syntactic dependency network is constructed by cumulative sentence structures, and the network is an emergent property of sentence structures (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005). Figure 2 shows an example of a small network containing 51 vertices.

The free software Pajek 1.22 was used for the network creation and computing (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

3. Comparison of WFN and LN

There are several statistical measures typically used for studying complex networks. This article makes observations about the following: an average degree, degree distribution, clustering coefficient, and average path length. The results of both WFN and LN are summarized in Table 1. The large sample sizes do not allow the use of the usual statistical tests (which were designed for much smaller amount of data; with sample sizes of tens of thousands almost all null hypotheses are rejected).

4 Of course, there is no definite agreement about direction among linguists. We are aware that subject–predicate relationship is reverse (subject governs predicate morphology), but on the other hand, we attach importance to valency. So, our approach is a compromise.

5 We used a simple graph that does not reflect a frequency of connections between particular nodes (as in a multigraph). This approach was used in all previous syntactic networks analysis and we follow it because of the possibility to compare the results.

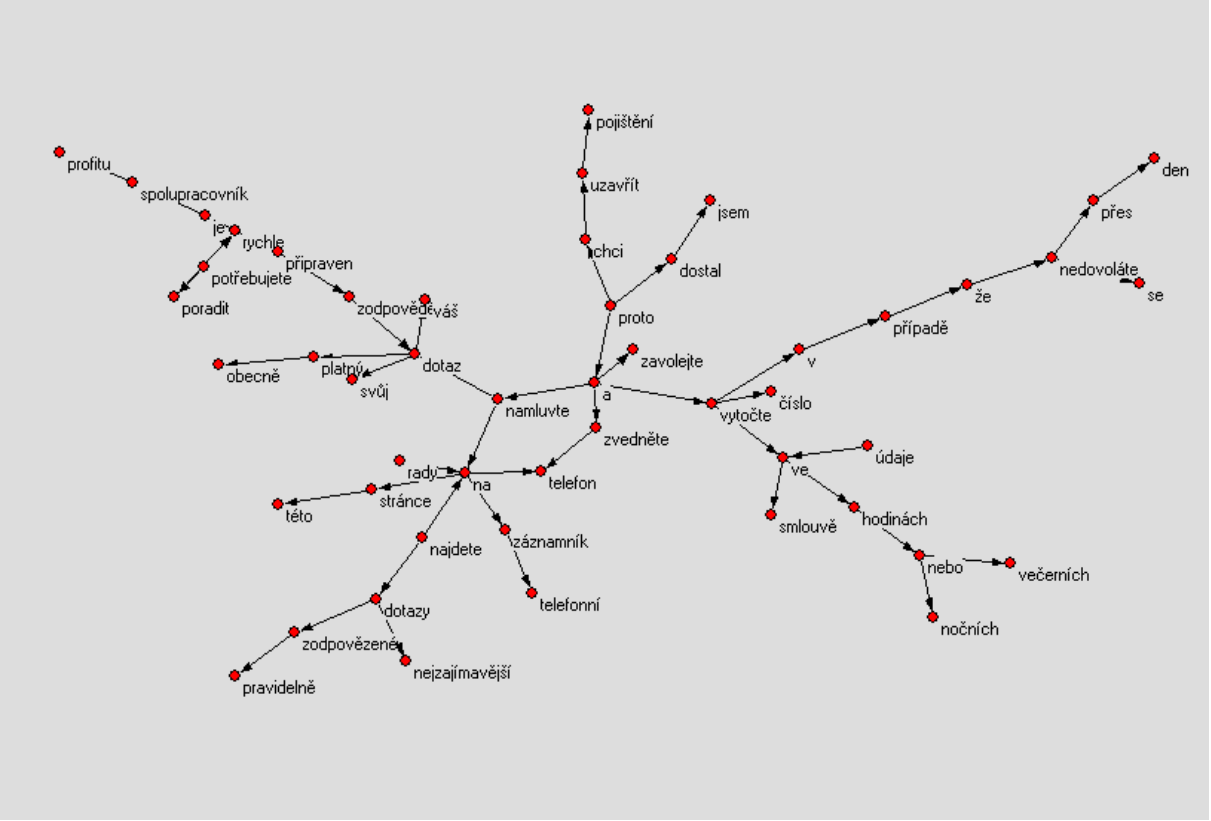


Figure 2. An example of a syntactic dependency network containing 51 vertices

Table 1
A summary of WFN and LN characteristics.

	WFN	LN
n	73989	36037
k	8.19	13.34
C	0.12	0.18
D	3.84	3.58

n = the number of vertices;

k = the average degree, it expresses the average connectivity (number of links) of all words;

C = the clustering coefficient, it is defined as the probability that two words that are neighbours of a given vertex are neighbours of each other;

D = the average minimum distance between words.

3.1 Average degree

The average degree expresses an average connectivity (number of links) of all words. It is given by

$$k = \Sigma_i k_i / N$$

where *k* is the number of undirected connections of each word *i* and *N* is the total number of words. The average degrees were measured on the undirected versions of both networks for simplicity reasons.

The comparison of WFN and LN (see Table 1) shows an average degree 1.6 times higher in LN. Furthermore, the network density, defined as

$$\delta = \frac{k}{n-1}$$

is about three times larger in LN (LN: δ = 0.000370185; WFN: δ = 0.000110692).

A discrepancy of *k* (and consequently the network density) between WFN and LN seems to be caused by inflection at first sight – if there are no inflected words in the language, the average degree of WFN and LN would be equal. But the influence of inflection upon average degree is not straightforward; syntactic relationships also have an important impact on it. Now, we will consider three possible types of syntactic connections between words with regard to inflection and the consequences for discrepancy size between WFN and LN (see Figure 3 and Table 2).

Type 1

There is only one word form (from all possible word forms) of a lemma connected to only one word form (from all possible word forms) of another lemma in WFN. This type has zero influence on discrepancy size, and it can appear in all three possible kinds of connections (with regard to inflection):

(i) between two indeclinable words (the only possible case);

(ii) between an indeclinable word and a declinable word which is connected to particular indeclinable word only by one word form;

(iii) between two declinable words, each of which is connected to the other only by one word form.

Type 2

There is only one word form (from all possible word forms) of a lemma connected to more than one word form of another lemma in WFN. This type causes a *higher average degree of WFN* and it can appear:

(i) between an indeclinable word and a declinable word which is connected to a particular indeclinable word by more than one word form;

(ii) between two declinable words, one of which is realized only by one word form (from all possible word forms) of a lemma and the other is connected to it by more than one word form.

Type 3

There is more than one word form of a lemma; each is connected to only one word form of different lemmas in WFN. This type causes a *higher average degree of LN* and it appears:

(i) between a declinable word which is realized by more than one word form and indeclinable words which are connected to only one word form of a declinable word (no indeclinable word can be connected to more than one word form of declinable word);

(ii) between a declinable word which is realized by more than one word form and

other declinable words; each word form of former declinable word is connected to a different word which is realized only by one word form (from all possible word forms) of a lemma.

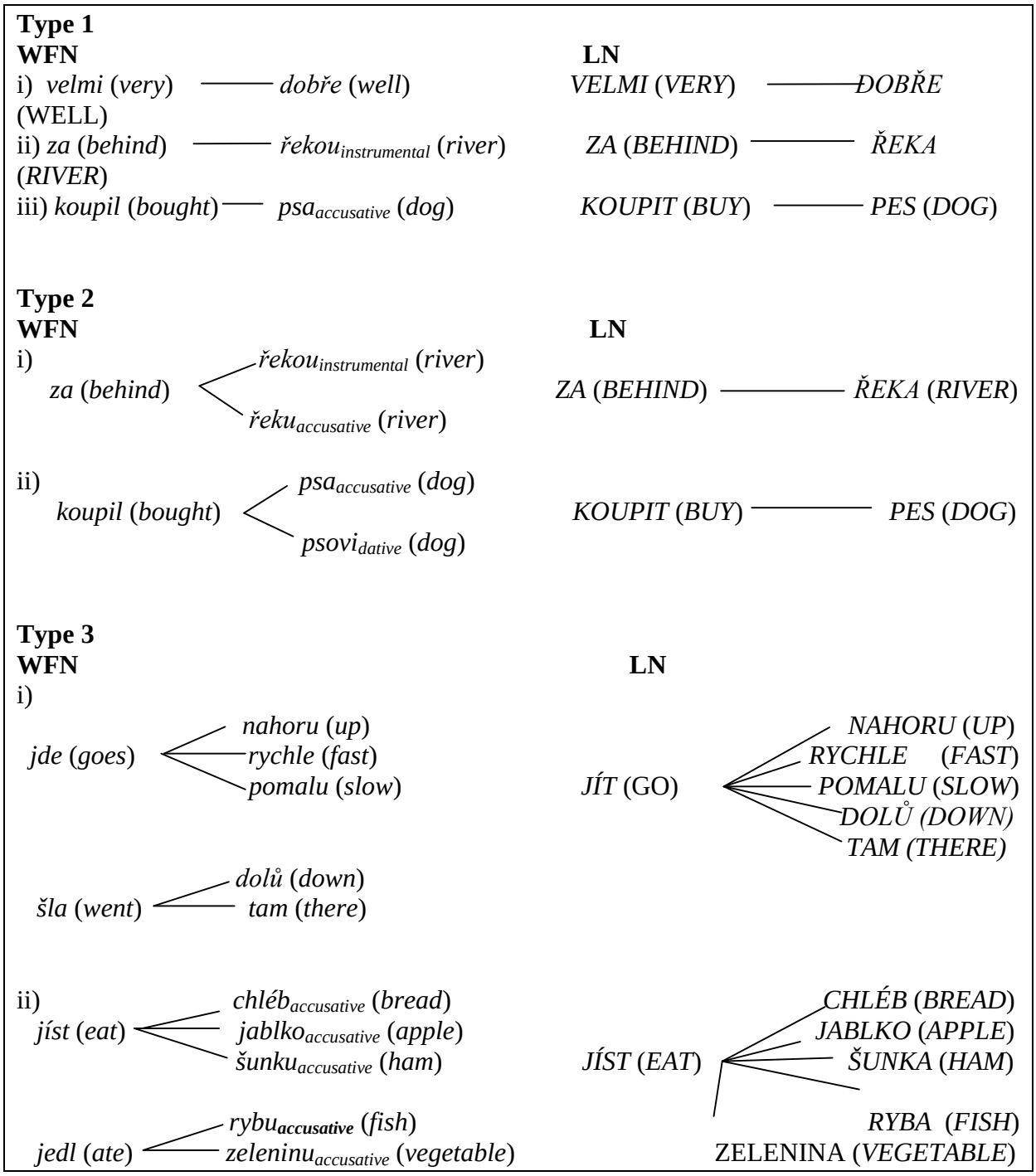


Figure 3. Three types of syntactic connections between words with regard to inflection. Each type is an illustrative example of a small language network. It shows how inflection and syntactic relationships determine a discrepancy size between WFN and LN. For values of *k* for each type see Table 2.

Table 2
Average degree values for each type of syntactic connections from Figure 3.

	WFN	LN
k_{type1}	1	1
k_{type2}	1.33	1
k_{type3}	1.42	1.67

The comparison of Types 1–3 shows that a discrepancy size is given by both inflectional richness (a number of particular word forms) and syntactic relationships. Moreover, an actual value of *k* is caused simultaneously by the grammar (if there is only one possible word form which could be connected to another word form e.g., connections between indeclinable words or a connection of an indefinite article with just a singular noun in English) and by the language usage (it determines how many word forms of a declinable word are realized in Types 2 and 3, and how many different indeclinable words are connected to a particular word form of a declinable lemma in Type 3). Consequently, a mere presence of inflection in the language does not necessarily entail higher values of average degree in WFN or LN and, theoretically, it is possible that the language with rich inflection will have the same average degree as the language with absolutely no inflection. So, the comparison of the average degree of WFN and LN cannot be used for typological characterisation of languages.

As for an observed discrepancy between WFN and LN (Table 1), the higher value of LN means that rich inflection is coupled with high lexical variability in actual language use. In other words, there is a strong tendency of a particular word form of a lemma to connect with many other lemmas through one word form of each of these lemmas (see Type 3). In Czech it is typical for instance of transitive verbs, as well as other types of verb, to have many raw word forms and which are connected to a host of different objects (accusative nouns). So again, the discrepancy is simultaneously a product of grammar and language usage.

3.2 Degree distribution

The degree distribution, $P(k)$, describes the number of nodes (e.g., word forms or lemmas) with a connectivity *k.* For most of large complex networks highly heterogeneous distribution is typical: the probability *P(k)* that a randomly chosen node in the network interacts with *k* other nodes decays as a power law, following $P(k) \sim k^{-\gamma}$ (Barabási, Albert 1999). Moreover, as Ferrer i Cancho (2005) shows, the distribution of word degrees could be a consequence of Zipf's law for word frequencies and the distribution, following the power law, should be a universal property of language.

However, in general our frequencies are not decreasing; in two cases we have $f(0) < f(1) > f(2) > f(3) > \ldots$. Hence we suggest the (shifted) gamma function as a model, namely,

$$f(x) = a(x+1)^b e^{-cx}.$$

We replaced the parameter *a* with the frequency $f(0)$. The obtained fit is very good (see Table 3) and, moreover, we do not reject the previous model; we generalize it (obviously, for $c = 0$ one obtains the power law).

Table 3
Fitting the gamma function to the degree distributions

	a	*b*	*c*	R^2
IN-distribution (WFN)	4665	11.936	6.100	0.9828
OUT-distribution (WFN)	28205	-0.440	0.322	0.9993
IN-distribution (LN)	1531	12.099	5.982	0.9653
OUT-distribution (LN)	13406	-0.211	0.444	0.9982

As can be seen, there are obvious differences between parameter values for nodes representing subordinate words and governing words.

We do not present particular frequencies in the degree distributions – that woud mean inserting four tables with approximately 250 lines.[6] The data can be sent upon request. A discrepancy between the previous model (i.e., power law distribution) and the result we obtained requires an explanation. As Newman (2006) shows, only "[f]ew real-world distributions follow a power law over their entire range, and in particular not for smaller values of the variable being measured. (...) [F]or any positive value of the exponent α the function $p(x) = Cx^{-a}$ diverges as $x \to 0$. In reality therefore, the distribution must deviate from the power-law form below some minimum value $x_{\min}$." In many analyses a distribution altogether below $x_{\min}$ is cut off and "one often hears it said that the distribution of such-and-such a quantity "has a power-law tail"." This precisely happened in our case. If we cut off nodes with frequency *f*(0), the all distributions follow the power law, so it seems reasonable to consider the power law as a universal language property for distributions for $x_{min} = 1$.

However, we consider the cutting off of distribution below some $x_{\min}$ improper in our analysis because the nodes with the frequency *f*(0) are linguistically important. They represent (1) word forms/lemmas which occur purely as terminal nodes of sentence structure, in the case of nodes which represent modifiers or (2) word forms/lemmas which occur purely as the highest elements in sentence structure (e.g., predicative verbs or various types of words in elliptical expressions), in the case of nodes which represent heads. That is why we propose the new model for degree distribution (see above).

Figure 4 presents cumulative in-degree and out-degree distributions in WFN and LN – the proportions of words whose input or output degree is *k* or more are shown. All graphs display a highly heterogeneous distribution; we can see that 90% of the words have less than 10 connections, whereas only 0.1% of all words have more than 100 connections. The distribution of words in LN follows the power law as well ($P(k) = (k+1)^{-b}$, see Table 4; we note that because of zero degrees the function had to be shifted to the left). Consequently, if the lemma network follows the power law (for $x_{min} = 1$), it means that the concept of lemma, despite of its artificiality, "obeys" the same language universal rule (Ferrer i Cancho 2005) and LN can be properly used for linguistic analysis.

[6] The data are available at the web page www.cechradek.ic.cz/@files/WFN_LN_distributions.xls

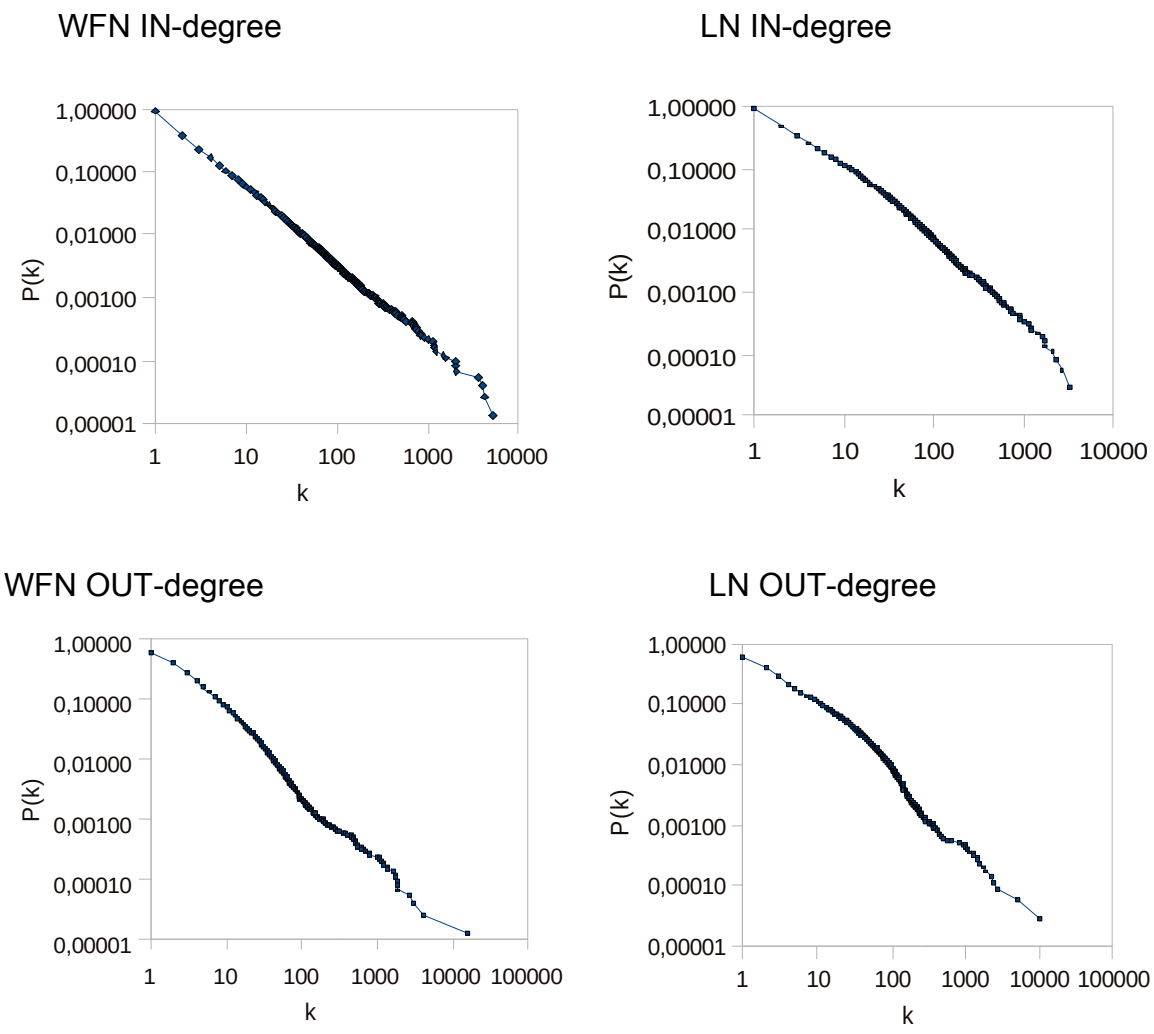


Figure 4. Cumulative degree distributions of network

Table 4
Fitting the power law function to the cumulative degree distributions.

	b	R^2
IN-distribution (WFN)	1.0170	0.8944
OUT-distribution (WFN)	1.0286	0.9766
IN-distribution (LN)	0.8721	0.9161
OUT-distribution (LN)	0.9211	0.9898

In this case one does not see any obvious differences in parameter values. We are aware of the vagueness of this statement; however, as we already mentioned, the huge amount of data is the reason why the usual statistical approaches cannot be applied.

3.3 Clustering coefficient and average path length

The clustering coefficient, *C*, expresses the probability that two nodes that are both neighbors of the same third node will be neighbors of one another (Newman 2001). The value of the clustering coefficient for a whole network is given by the average over all nodes and it indicates interconnectivity of a complex network. Because of problems with a clustering coefficient measurement in directed networks (Caldarelli 2007), values of *C* were measured on the undirected versions of both networks, WFN and LN.

Table 1 shows that there is a higher clustering coefficient in LN than in WFN. As in average degree (Section 3.1), a discrepancy is caused by the presence of declinable words in the language and by syntactic relationships. Next we will consider three types of syntactic relationship with regard to inflection and consequences for differences of *C* between WFN and LN.

Type 1

There is an indeclinable word connected to only indeclinable words. Obviously, interconnectivity among indeclinable words has zero influence on differences between clustering coefficients of WFN and LN, because all syntactic relationships are the same in WFN and LN thanks to the absence of inflection.

Type 2

There is an indeclinable word connected to a declinable word or words. Theoretically, all three possible cases with regard to discrepancies of C between WFN and LN should appear:

(i) higher *C* of LN,
(ii) equal C of WFN and LN,
(iii) higher C of WFN.

But if we take account of grammar properties of the observed language, we would be able to hypothesize, at least partly, what kind of a clustering coefficient discrepancy (i.e., equal *C* of WFN and LN, higher *C* of WFN or higher *C* of LN) between WFN and LN should be more often detected than the others.

For (i) let us start with the clustering coefficient higher in LN than in WFN. For illustration (see Figure 5), assume the connections between the preposition, e.g., $o_{(locative)}$ (*about*)[7], and two nouns, e.g., *stůl* (*table*) and *noha* (*foot*). The preposition $o_{(locative)}$ (*about*) governs the grammatical case of dependent nouns, but there is no possible syntactic relationship between the two nouns in the locative case in Czech. Consequently, the clustering coefficient of preposition $o_{(locative)}$ *(about)* equals zero in WFN. Now, assume the same example in LN: all word forms of declinable words fall to one lemma, so if there is a possible syntactic relationship between *stůl* (*table*) and *noha* (*foot*) in other grammatical cases i.e., in an attributive connection $noha_{nominative}$ $stolu_{genitive}$ (*table foot*), there is a potentiality that lemmas *O* (*ABOUT*), *STŮL* (*TABLE*), and *NOHA* (*FOOT*) can be interconnected in LN. So, if there is an actual connection (it depends on the usage) between these nouns, the clustering coefficient is higher in LN than in WFN (Figure 5).

7 Preposition *o* can also be used with accusative noun in Czech; in presented example we consider only locative coligation.

.

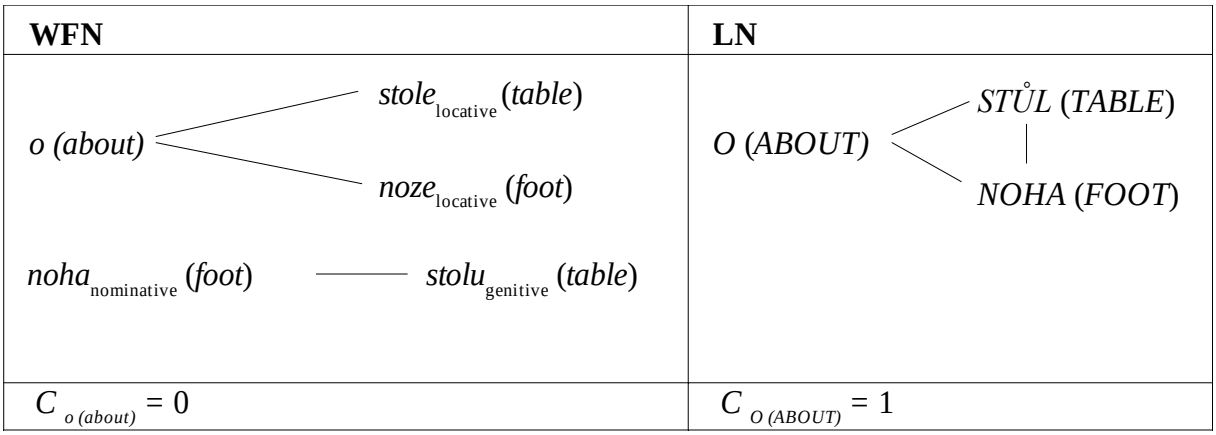


Figure 5. An example of small networks, WFN and LN, based on five raw word forms. The connection between declinable words *noha*$_{nominative}$ *stolu*$_{genitive}$ (*foot table*) in WFN causes higher C of lemma *O* (*ABOUT*) in LN

For (ii), now, consider an instance when clustering coefficients in WFN and LN are equal. Again for illustration (see Figure 6), assume the connections between genitive preposition *bez* (*without*) and two nouns, *stůl* (*table*) and *noha* (*foot*). Contrary to nouns in locative, genitive nouns could connect each other e.g., *nohy*$_{genitive}$ *stolu*$_{genitive}$ (*table foot*). So if these nouns are actually connected in WFN, the clustering coefficient is equal in WFN and LN.

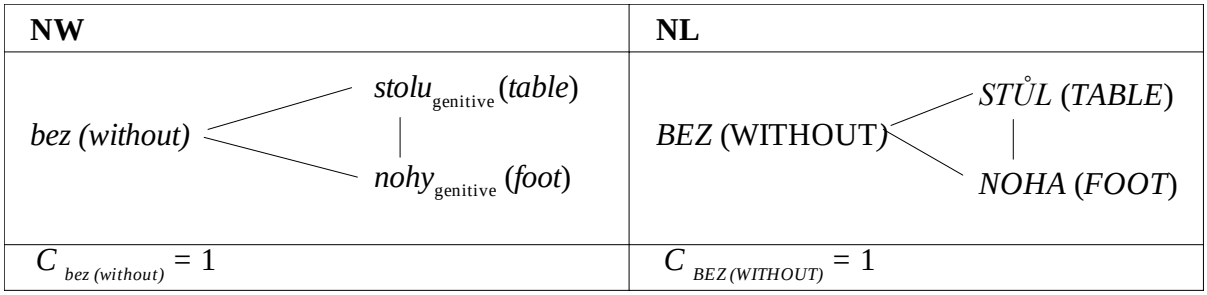


Figure 6. An example of small networks based on three raw word forms. The connection between dependent genitive nouns, *nohy*$_{genitive}$ *stolu*$_{genitive}$ (*table foot*), causes equality of C of preposition *bez* (*without*) in WFN and LN.

Of course, an equality of clustering coefficients in WFN and LN would be also given, if there is no actual connection between declinable words in WFN and LN – a value of the clustering coefficient of an indeclinable word equals zero in both networks.

Finally, for (iii), the higher value of the clustering coefficient in WFN would be possible, if there is a connection between indeclinable word and at least two word forms of the same lemma of declinable word, and moreover between these same word forms in WFN: e.g., all instances as *bez písně*$_{genitive\ sg.}$ (*without song*), *bez písní*$_{genitive\ pl.}$ (*without songs*), and *písně*$_{genitive\ sg.}$ *písní*$_{genitive\ pl.}$ (*song of songs*) have to be present in WFN for the higher value of the clustering coefficient in WFN (see Figure 7).

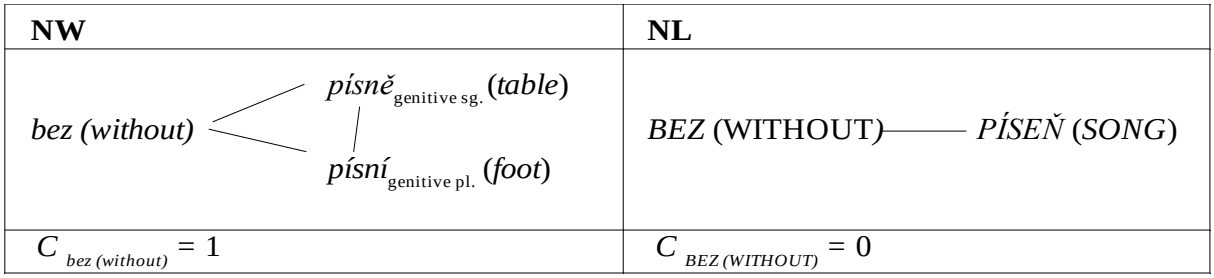


Figure 7. An example of small networks based on three raw word forms. The connection between two dependent word forms of the same lemma, *PÍSEŇ* (*SONG*), causes higher *C* of preposition *bez* (*without*) in WFN.

Because case (i) is more typical than case (ii), and case (iii) is very rare in Czech we can expect that words with no inflection should have more often a higher clustering coefficient in LN than in WFN. So, with regard to syntactic connections between indeclinable and declinable words the higher *C* in LN (Table 1) is not surprising.

Type 3

Type 3 represents syntactic relationships between declinable words. All the three possible cases with regard to discrepancies of *C* between WFN and LN should appear: equal *C* of WFN and LN, higher *C* of WFN or higher *C* of LN. Contrary to Type 2, it is primarily the language usage that determines how many word forms of the lemma are realized and which are connected to each other in WFN, so the influence of grammar is much weaker than in Type 2. Consequently, it is practically impossible to hypothesize which kind of clustering coefficient discrepancy (i.e., equal *C* of WFN and LN, higher *C* of WFN or higher *C* of LN) between WFN and LN should be more often detected than the others.

Moreover, the clustering coefficient is crucial in the characterization of complex networks: a high *C* in real networks, in comparison of *C* in Erdős-Rényi random networks, indicates the small world structure of real networks (Watts, Strogatz, 1998). Table 3 shows that $C_{WFN} >> C_{random}$ and $C_{LN} >> C_{random}$. The small world phenomenon means that despite the large number of network elements (e.g., word forms or lemmas) the distance between them is strikingly small. The value of this distance is called average path length, *D*, and it expresses the shortest distance (it is defined by a number of links) between any randomly chosen pair of nodes of a network. Values of *D* in both WFN and LN are close to D_{random} (see Table 5) which expresses the value of *D* for Erdős-Rényi random network. It exhibits, with considerably high clustering formation that both networks are really small worlds (Watts, Strogatz, 1998).

Table 5
Clustering coefficients and average path lengths of real and random networks.

	C	*D*
WFN	0.12	3.84
WFN_{random}	0.0009	5.57
LN	0.18	3.58
LN_{random}	0.004	4.35

4. Discussion and future work

We presented a study which compares two syntactic dependency networks based on the same language data; in one network each node represents a raw word form (WFN), in the other network each node represents a basic word form, lemma (LN). The analysis shows discrepancies between WFN and LN in observed values. These discrepancies are caused by the inflectional characteristics of the Czech grammar, syntactic relationships, and by language usage.

As regards *average degree* we can partly hypothesize a relationship between the typological character of language and the average degree discrepancy between WFN and LN:

- networks based on languages with no inflection (as a highly isolating language) will have zero discrepancy,
- networks based on languages with low inflection (as English) will have zero discrepancy or higher average degree of WFN,
- for networks based on highly inflectional languages it is not possible to make theoretical hypotheses; all the three potential kinds of discrepancy could appear because the discrepancy value is significantly influenced by language usage. The present study shows why one could expect a higher average degree in LN (Section 3.1), but only further observations of networks based on different highly inflectional languages could reveal typical characteristics of WFN and LN with regard to average degree.

The kind of *clustering coefficient* differences between WFN and LN could be partly hypothesized (for words with no inflection) by closer grammar observation, as we presented in Section 3.3. Not only grammar or typological characteristics play a crucial role for clustering coefficient discrepancy, language usage also notably influences the clustering coefficient as well the average degree.

The findings presented in the article are important for potential matching of individual lemma networks based on different languages; for appropriate comparison of these networks one has to take into account at least typological characteristics of languages. Because of the importance of language usage one should also consider the possible influence of text types, but only further observations of networks based on different registers could reveal a real impact of text types on both the average degree and clustering coefficient.

Acknowledgements

R. Čech and J. Mačutek were supported by GAČR (Czech Science Foundation) No. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach) and by the Lise Meitner Stipendium (FWF, Austria), respectively.

References

Barabási, A.L., Albert, R. (1999). Emergence of Scaling Random Networks. *Science 286, 509-512.*

Caldarelli, G. (2007). *Scale-Free Networks: Complex Webs in Nature and Technology.* Oxford: Oxford University Press.

Caldeira, S.M.G., Lobão, T.P., Andrade, R.F.S., Neme, A., Miranda, J.G.V. (2006). The network of concepts in written texts. *European Physical Journal, B 49, 523-529.*

Dorogovtsev, S.N., Mendes, J.F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B 268, 2603–2606.*

Ferrer i Cancho, R. (2005). The structure of syntactic dependency networks: insight from recent advances in network theory. In: Altmann, G., Levickij, V.V., Perebyinis, V.

(eds.), *Problems of Quantitative Linguistics: 60-75.* Chernivtsi: Ruta.

Ferrer i Cancho, R. (2006a). Why do syntactic links not cross? *Europhysics Letters 76, 1228-1235.*

Ferrer i Cancho, R. (2006b). When language breaks into pieces. A conflict between communication through isolated signals and language. *Biosystems 84, 242–253.*

Ferrer i Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems 11 (3), 394–414.*

Ferrer i Cancho, R., Riordan, O., Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London Series B 272, 561–565.*

Ferrer i Cancho, R., Solé, R.V. (2001). The small-world of human language. *Proceedings of the Royal Society of London B 268, 2261–2265*.

Ferrer i Cancho, R., Solé, R.V., Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E 69, 051915.*

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.

Holanda, A.J., Pisa, I.T., Kinouchi, O., Martinez, A.S., Ruiz, E.E.S. (2004). Thesaurus as a complex network. *Physica A 344, 530–536*.

Hudson, R. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Ke, J., Yao, Y. (2008). Analysing Language Development from a Network Approach. *Journal of Quantitative Linguistics 15(1), 70–99.*

Liu, H. (2008). The complexity of Chinese syntactic dependency networks. *Physica A 387, 3048–3058.*

Mehler, A. (2007). Large Text Networks as an Object of Corpus Linguistic Studies. In: Lüdeling, A., Kytö, M. (eds.), *Corpus Linguistic. An International Handbook: 328-382*. Berlin: Mouton de Gruyter.

Motter, A.E., de Moura, A.P.S., Lai, Y.-Ch., Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E 65, 065102(R).*

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E 64, 025102(R).*

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics 46, 323–351.*

Ninio, A. (2006). *Language and the learning curve. A new theory of syntactic development.* Oxford: Oxford University Press.

Sigman M., Cecchi, G.A. (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America 99(3), 1742-1747*.

Soares, M., Corso, G., Lucena, L.S. (2005). The network of syllables in Portuguese. *Physica A, 355 (2-4), 678–684.*

Solé, R.V. (2005). Syntax for free? *Nature 434, 289.*

Steyvers, M., Tenenbaum, J.B. (2005). The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science 29(1), 41–78.*

Watts, D.J., Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature 393, 440–442.*

Glottometrics 19, 2009, 99-101

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.

XLI. William Palin Elderton (1877-1962)

Geb. 26.6.1877 in Fulham, gest. 6.4.1962. Elderton absolvierte die *Merchant Taylor's School*, konnte aber wegen des frühen Todes seines Vaters, William Alexander Elderton, kein Studium absolvieren und begann stattdessen sein Berufsleben mit 17 Jahren bei der *Guardian Assurance Company*. Die Weiterentwicklung statistischer Verfahren und ihre Anwendung vor allem im Bereich des Versicherungswesens, aber darüber hinaus in vielen anderen Feldern, prägten sein Leben. In diesem Zusammenhang werden Eldertons langjährige Kontakte mit Karl Pearson immer wieder hervorgehoben. Er nahm viele leitende Positionen in beiden Bereichen ein und erhielt neben weiteren Auszeichnungen 1938 den Doktorgrad ehrenhalber von der Universität Oslo verliehen. Er war über Jahrzehnte Mitglied des *University of London Board of Studies in Statistics* und des *Council* der *Royal Statistical Society*. In beiden Weltkriegen stellte er seine Kenntnisse in der Statistik der Regierung, speziell dem *Ministry of Shipping*, zur Verfügung.

Es ist hier nicht der Ort, all die Verdienste Eldertons aufzuzählen; dazu sei auf die Nachrufe verwiesen. Menzler (1962: 671) betont: „He could talk with interest and knowledge about literature, philosophy and architecture, particularly old churches." Hier kommt auch seine Bedeutung für die Quantitative Linguistik ins Spiel: So ist Elderton (1949) etwa gleichzeitig mit Čebanov (1947) und noch vor Fucks (1955) der zweite Autor, der für die Verteilungen von Wortlängen in Texten ein mathematisches Modell, in seinem Fall die geometrische Verteilung, vorschlug. Čebanov (1947) hatte dagegen die Poisson-Verteilung als ein geeignetes Modell erkannt. Auf Eldertons Daten griff Herdan (1960: 183f., 187; 1966: 285f.) zurück. Der bedeutsame Artikel (Elderton 1949) ist zwar in Köhlers Bibliographie aufgeführt (Köhler 1995), findet jedoch im Handbuch von Köhler, Altmann & Piotrowski (2005) keine Erwähnung, auch nicht im einschlägigen Beitrag Best (2005). Um so verdienstvoller ist es, dass Grybek (2006: 19-23) diese Arbeit wieder aufgreift, eingehend behandelt und in die Entwicklung der Wortlängenforschung einordnet. In Best (2009) schließlich wird nachgeholt, was längst hätte geschehen sollen: Es wird gezeigt, dass man für Eldertons Daten zu Wortlängen in sehr unterschiedlichen Texten, Textsorten und -korpora Modelle finden kann, die sich aus der Theorie von Wimmer u.a. (1994) und Wimmer & Altmann (1996) entwickeln lassen und die sich in vielen anderen Untersuchungen zum Englischen bereits bewährt haben (vgl. dazu die Hinweise in Best 2009). Es ergab sich, dass für die Gedichte die Poisson-Verteilung und für alle anderen Texte die positive Singh-Poisson-Verteilung geeignet sind.

Der Beitrag hat allerdings eine Vorgeschichte. Elderton (1949: 136) berichtet, dass er die Idee, Wortlängen könnten sich der geometrischen Verteilung fügen, im März 1946 von seinem Freund (Prof. A.C. Aitken, F.R.S.) vermittelt von J. B. Molony übernommen habe, der Wortlängen in Fitzgeralds Übersetzung von Omar Khayyám, *Rubáiyát* ausgezählt hatte und

eine Ähnlichkeit mit der geometrischen Reihe feststellte.[1] Die Idee, die geometrische Reihe könne ein gutes Modell sein, fiel bei Elderton auf fruchtbaren Boden, weil er schon vorher in ganz anderen Feldern auf dieses Modell gestoßen war. Dieser Anregung ist Elderton also nachgegangen und hat laut eigenem Bekunden am 17.3.1947 einen Vortrag vor dem *Insurance Institute of Norwich* über *Cricket Scores, Fire and Accident Claims and Gray's Elegy* gehalten und Ergebnisse vorgetragen.

Man kann damit feststellen, dass die Versuche, mathematische Modelle für die Wortlängenverteilung in Texten zu entwickeln, etwa gleichzeitig in England (Molony/Elderton) und der Sowjetunion (Čebanov) verfolgt wurden, wobei unterschiedliche Modelle ins Auge gefasst wurden. Knapp ein Jahrzehnt später entwickelte Fucks (1955, 1956) die gleiche Idee wie Čebanov.

Werke Eldertons (Auswahl)

1906 *Frequency-Curves and Correlation.* London.

1914 & Richard C. Fippard. *The Construction of Mortality and Sickness Tables: a Primer.* London: Black.

1920 & Ethel M. Elderton. *Primer of statistics.* London: Black.

1924 *Krivyja raspredelenija čislennostej i korreljacija.* Moskva: Izd. Centr. statistič. upravlenija.

1924 *The Mortality of Annuitants 1900-1920: Investigation and Tables.* London: Layton.

1927 *Frequency-Curves and Correlation.* 2. ed. London: C. E. Layton.

1928 *Shipping problems. 1916-1921.* London: Black.

1931-34 William Morgan, F.R.S. (1750-1833). *Transactions of the Faculty of Actuaries 14, S. 1-20.*

1938 *Frequency-Curves and Correlation.* 3d ed. Cambridge: The University Press.

1938 *The Impossibility of War Risk Insurance: a Paper Read before the Insurance Institute of London on 15th March, 1938.* Cambridge: University Press.

1943 The Mortality of Adult Males since the Middle of the Eighteenth Century as Shown by the Experience of Life Assurance. In: *Journal of the Royal Statistical Society, Bd. 106 (1943), S. 1-31.*

1947 Merchant Seamen During the War. In: *Journal of the Institute of Actuaries, Bd. 73 (1947), 2, S. 250-284.*

1949 A Few Statistics on the Length of English Words. *Journal of the Royal Statistical Society, Series A (General), Vol. CXII, Part IV, S. 436-445.*

1953 *Frequency-Curves and Correlation.* 4. ed. Cambridge: University Press 1953.

1969 & Johnson, Norman Lloyd (1969). *System of Frequency Curves.* Cambridge: University Press.

In dieser Liste sind nur wenige der Werke Eldertons enthalten, vor allem seine Monographien und der Aufsatz von 1949, der Anlass für diesen Artikel gibt. Tappenden (1962: 248, passim) bezeugt eine Fülle von Publikationen, die hier jedoch – weil für die Quantitative Linguistik fachfremd – nicht alle dokumentiert werden müssen. Die obige Liste soll nur einen Eindruck von seinem Schaffen vermitteln.[2]

[1] Daten zu diesem Text liefert Herdan (1960: 184) mit Berufung auf Elderton (1949: 136), wo diese Daten sich jedoch nicht finden.

[2] Unter der Adresse http://www.actuaries.org.uk/knowledge/publications/jia_tfa kann man viele der Arbeiten Eldertons leicht finden.

Nachrufe

G., R.Ll. (1962-1964). The Late Sir WILLIAM PALIN ELDERTON K.B.E., Ph.D. (Oslo). *Transactions of the Faculty of Actuaries 28, 193-195.*

Menzler, F.A.A. (1962). Sir William Palin Elderton, 1877-1962. *Journal of the Royal Statistical Society, Series A (General), Vol. 125, No. 4, 669-672.*

Pearson, E.S. (1962). William Palin Elderton (1877-1962). *Biometrika 49, 297-303.* (Auf S. 296 findet sich ein Portrait von Elderton.)

Tappenden, H.J. (1962). Sir William Palin Elderton, K.B.E., Ph.D. (Oslo). *Journal of the Institute of Actuaries 88, 245-251.*

Literatur

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

Best, Karl-Heinz (2009). Wortlängen im Englischen. In Arbeit.

Čebanov, Sergej Grigor‘evič (1947). O podčinenii rečevych ukladov ‘indoevropejskoj’ gruppy zakonu Puassona. *Doklady Akademii Nauk SSSR. Tom 55/2, 103-106.*

Fucks, Wilhelm (1955). Theorie der Wortbildung. *Mathematisch-Physikalische Semesterberichte. Bd. 4, 195-212.*

Fucks, Wilhelm (1956). Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen. *Nachrichtentechnische Forschungsberichte 3, 7-21.*

Grzybek, Peter (2006). History and Methodology of Word Length Studies. The State of the Art. In: Grzybek, Peter (ed.). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues* (S. 15-90). Dordrecht: Springer.

Herdan, Gustav (1960). *Type-Token Mathematics. A Textbook of Mathematical Linguistics.* ‘s-Gravenhage: Mouton.

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

Köhler, Reinhard (1995). *Bibliography of quantitative linguistics*. Amsterdam: John Benjamins.

Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.) (2005), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics An International Handbook.* Berlin/ N.Y.: de Gruyter.

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

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

Karl-Heinz Best, Göttingen