Glottometrics 18
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. 18 (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 18 (2009) **ISSN 2625-8226**

Contents

Glottometrics 18, 2009, 1-12

Distribution of complexities in the Vai script

Andrij Rovenchak[1]*, Lviv*
Ján Mačutek[2]*, Bratislava*
Charles Riley[3]*, New Haven, Connecticut*

Abstract. In the paper, we analyze the distribution of complexities in the Vai script, an indigenous syllabic writing system from Liberia. It is found that the uniformity hypothesis for complexities fails for this script. The models using Poisson distribution for the number of components and hyper-Poisson distribution for connections provide good fits in the case of the Vai script.

Keywords: Vai script, syllabary, script analysis, complexity.

1. Introduction

Our study concentrates mainly on the complexity of the Vai script. We use the composition method suggested by Altmann (2004). It has some drawbacks (e. g., as mentioned by Köhler 2008, letter components are not weighted by their lengths, hence a short straight line in the letter G contributes to the letter complexity by 2 points, the same score is attributed to each of four longer lines of the letter M), but they are overshadowed by several important advantages (it is applicable to all scripts, it can be done relatively easily without a special software). And, of course, there is no perfect method in empirical science. Some alternative methods are mentioned in Altmann (2008).

Applying the Altmann's composition method, a letter is decomposed into its components (points with complexity 1, straight lines with complexity 2, arches not exceeding 180 degrees with complexity 3, filled areas[4] with complexity 2) and connections (continuous with complexity 1, crisp with complexity 2, crossing with complexity 3). Then, the letter complexity is the sum of its components and connections complexities. E. g., the letter O is assigned complexity 8 (2 arches, 2 continuous connections), the letter X has complexity 7 (2 straight lines, 1 crossing). See Altmann (2004) and Mačutek (2008) for a more detailed discussion on the method. In some cases the method is not unambiguous, e. g., sometimes a researcher must decide if he considers a thick line or a filled area with its contours. Different fonts of the same script usually yield different complexities.

From among scripts so far analyzed with respect to complexity we mention Latin (fonts Arial and Courier New, Altmann 2004), Cyrillic (its Ukrainian version, Buk, Mačutek and Rovenchak 2008) and several types of runes (Mačutek 2008).

[1] Department for Theoretical Physics, Ivan Franko National University of Lviv, 12 Drahomanov St., Lviv, UA-79005, Ukraine, e-mail: andrij@ktf.franko.lviv.ua, andrij.rovenchak@gmail.com
[2] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 824 48 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com
[3] Sterling Memorial Library, Yale University, 120 High Street, New Haven, CT 06511, USA, e-mail: charles.riley@yale.edu
[4] Altmann (2004) proposed complexity 1 for filled areas.

2. The geographic range of the Vai people and their language

Vai (also Vei, Vy, Gallinas, Gallines, phonetically [vaɪ]) is a Western Mande language belonging to the Niger-Congo language macrofamily. It is spoken by some 144,000 people, of which about 122,000 live in Liberia and some 22,000 in Sierra Leone[5]. The territory of the Vai speakers is shown in the map (Fig. 1). It is located on the Atlantic coast and stretches from the Lake Mabesi in the West to the Lofa River in the East, its northern boundary lying some ten miles south from the city of Potoru in Sierra Leone (roughly, this territory lies between 11°40′W and 11°00′W, and below 7°20′N). Note, that the Sierra Leone part is to a large extent shared with other peoples, namely, Mende (belonging to the Mande family) and Gola (speaking a language from the Atlantic family).

Figure 1. Map showing the location of the Vai country (dark shaded area).

The Vai language has a remarkable phonology, reflecting its lexical history (Welmers, 1976). There are seven oral vowels [a, ɛ, e, i, ɔ, o, u], five nasal vowels[6] [ã, ɛ̃, ĩ, ɔ̃, ũ], and 31 consonant [b, ɓ, ʧ, d, ɗ, f, g, gb, h, ʤ, k, kp, l, m, mɓ, mgb, n, nɗ, ɲ, ɲʤ, ŋ, ŋg, p, s, t, v, w, j, z, r, ʃ]. The consonants [r, ʃ] are found only in relatively recent loans (John Singler, personal communication, 2005), and [ʧ] is probably relatively recent as well (Ofri-Scheps 1991).

[5] There is a large uncertainty about these numbers. The most recent available data can be found in Gordon (2005). These are 89,500 in Liberia and 15,500 in Sierra Leone for 1991. The results of recently conducted 2008 Census in Liberia with regard to ethnicity are not known yet. To make the estimation of the population, we took the population growth in Liberia between 1984 and 2008 (Census 2008). We are grateful to Prof. William Kory for the discussion on this issue. It must be noted that the number of the Vais in Sierra Leone defined by primary language use during the 2004 Census is significantly lower, about 2,500 (Thekeka Conteh, personal communication on July 10, 2008), when comparing to the estimated number of 22,000 obtained from the population growth in Sierra Leone.

[6] Note that the status of [ũ] is unclear: it is represented by a single syllabic sign hũ, probably an obsolete one (Priest 2004) and not included in the vowels list by modern authors (Welmers 1976, Ofri-Scheps 1991). Ofri-Scheps excludes [ĩ] as well, which is elsewhere represented by a single sign hĩ.

3. Vai script: organization, history, usage

The Vai script was created in 1820–30s in Jondu, Cape Mount, Liberia. Traditionally, Mɔmɔlu Duwalu Bukɛlɛ (?–1850) is credited as an inventor of the script. He presumably was assisted in this work by his five friends (Dalby 1967). It appears possible that the emergence of the Vai script is linked with the "stimulus diffusion" from the Cherokee syllabic script. Despite large distance from the Cherokees living in Oklahoma and the Vais in Liberia, this writing system could be known from the American mission in early 1820s: one of the missionaries was Austin Curtis, the Cherokee himself and at the same time one of the leading men in the Vai country (Tuchscherer & Hair 2002; Mafundikwa 2004; Tuchscherer 2005; 2007). The Vai script was widely used in the private sphere, where it survived until today. According to some data, every fifth Vai man can write it (Singler 1983; Tuchscherer 2005; 2007). The translations of Qur'an and Bible in the Vai script are also known. It is assumed that the Vai script became a direct stimulus for the creation of several other indigenous writing systems in the Western Africa, in particular Mende, Loma, Kpelle (Dalby 1967), and Bambara (Galtier 1987).

The Vai script is a syllabary. Not all the combination of the above-listed sounds are equally possible, thus the total number of syllables is far below the available arithmetically calculated maximum of some 400. The Vai syllabary counts over 200 signs, of which seven stand for individual oral vowels, two can be treated as independent nasals [ã, ɛ̃], remaining signs denoting open syllables plus one sign for syllabic [ŋ] (Dalby 1967; Jensen 1969; Coulmas 2004), see Table 1. Since the creation, many Vai signs changed their appearance, and the syllabary was standardized in 1899 and finally in 1962 by a committee in the University of Liberia (Singler 1996: 594). It is worth to note that in 1911 Momolu Massaquoi made an attempt to fill in the gaps of the original syllabary modifying the signs by adding diacritical marks (Massaquoi 1911). Some Massaquoi additions survived until today, in particular his signs for r- and ʃ-series; some were abandoned by Vai practitioners.

While Vai is a tonal language, the tones are not marked in the script. Such deficiency is however common even for many Roman orthographies used in African languages (Bird 1999). Of indigenous African writing systems, only in Bassa are the tones marked in a systematic way (Coulmas 2004), and – with a much higher precision – in the N'ko alphabet (Vydrine 1999).

In fact, "the basic unit of the system is more accurately the mora" (Singler 1996: 594). Syllables containing a long vowel are written with two signs, the second one often belonging to the h-series. The velar nasal [ŋ] is treated as a separate unit. Surprisingly, similar situation with mora-to-sign correspondence is found in, e. g., Japanese *kana*.

The Vais prefer European punctuation and digits, however, there exist native Vai punctuation signs (Jensen 1969, Massaquoi 1911). The Vai digits are also known but they are not widely used (Everson et al. 2006), these are modified European digits to conform the Vai style.

The direction of the Vai script is from left to right in horizontal lines going from top to bottom.

4. Some quantitative analyses

Table 1 contains the list of the Vai syllabic signs based mainly on the Dukor typeface (courtesy of Evertype). This font reflects the style given in Tucker (1999). Recently, the Vai script

became a part of the Unicode standard, version 5.1. This fact also helps to standardize the shape of individual signs, which varies a lot in handwritten texts.

Table 1
Complexity of Vai letters

	sign	transliteration	components	connections	complexity
1		a	4×1+1×2+2×3	2×1+1×2	16
2		ɛ	6×3	5×1+3×3	32
3		e	3×2+4×3	4×1+2×2	26
4		i	4×2+2×3	5×2	24
5		ɔ	6×3	5×2	28
6		o	6×2	5×2	22
7		u	2×1+3×2+2×3	5×2	24
8		ã	3×3	2×1+1×3	14
9		ɛ̃	1×2+5×3	2×1+4×2	27
10		ba	4×3	3×1+2×3	21
11		bɛ	2×1+4×2	3×2	16
12		be	2×1+2×2+6×3	6×1+3×2	36
13		bi	6×3	4×1+1×2+2×3	30
14		bɔ	1×2+8×3	6×1+2×2+2×3	42
15		bo	2×1+2×2+4×3	3×1+2×2	25
16		bu	6×3	6×1	24
17		ɓa	1×1+4×2+1×3	6×2	24
18		ɓɛ	1×2+1×3	—	5
19		ɓe	2×2+1×3	2×2	11
20		ɓi	2×2+4×3	4×1+4×2	28
21		ɓɔ	3×2+1×3+1×**2***	4×2	19
22		ɓo	1×1+5×2	4×2	19
23		ɓu	5×3	4×1+5×2	25
24		ʧa	2×2+2×3	1×1+3×2	17
25		ʧɛ	2×2+4×3	2×1+3×2	24
26		ʧe	5×3+2×**2**	4×1+2×3	29
27		ʧi	2×2+2×3	1×1+3×2	17
28		ʧɔ	2×1+2×2+2×3	6×2	24
29		ʧo	4×1+1×2	—	6
30		ʧu	2×1+2×2+1×3	1×2+1×3	14
31		da	2×1+1×2+2×3	2×2	14
32		dɛ	4×2+4×3	4×1+4×2+4×3	46
33		de	7×3	2×1+4×2+1×3	34
34		di	1×2+3×3	2×1+2×2	17

	sign	transliteration	components	connections	complexity
35		dɔ	2×1+4×2	3×2	16
36		do	4×2+3×3	8×2	33
37		du	4×2	2×2+1×3	15
38		ɗa	4×2+1×3	6×2	23
39		ɗɛ	2×1+3×2	—	8
40		ɗe	4×2	2×2	12
41		ɗi	2×1+2×3+1×**2**	2×1	12
42		ɗɔ	2×2+3×3	2×1+2×2+1×3	22
43		ɗo	6×2	4×2+1×3	23
44		ɗu	2×1+1×2+1×3	1×2	9
45		fa	1×2+4×3	2×1+2×2	20
46		fɛ	5×2	4×2	18
47		fe	2×2+2×3	2×1+2×2	16
48		fi	4×3	3×2	18
49		fɔ	1×2+4×3	4×1+1×2+2×3	26
50		fo	2×2+2×3	3×2+1×3	19
51		fu	1×2+6×3	5×1+2×2+1×3	32
52		ga	6×3	8×2	34
53		gɛ	3×2+2×3	4×2+1×3	23
54		ge	3×2	2×3	12
55		gi	2×1+1×2+4×3	1×1+3×2+1×3	26
56		gɔ	3×1+2×3	2×1	11
57		go	4×2	2×2+1×3	15
58		gu	1×1+2×2+2×3	2×1+2×2	17
59		gɛ̃	4×2+2×3	6×2	26
60		gba	1×2+2×3	5×2	18
61		gbɛ	4×2	4×2+1×3	19
62		gbe	2×2	1×2	6
63		gbi	4×1+2×2	1×3	11
64		gbɔ	6×2	9×2	30
65		gbo	4×2	4×2	16
66		gbu	2×2+6×3	4×1+6×2	38
67		gbɛ̃	2×1+4×2	4×2+1×3	21
68		gbɔ̃	9×2	9×2	36
69		ha	2×2+3×3	6×2	25
70		hɛ	8×3	7×1+4×3	43
71		he	1×2+3×3	2×2+1×3	18
72		hi	5×2+2×3	5×2+1×3	29
73		hɔ	2×2+6×3	7×2	36

	sign	transliteration	components	connections	complexity
74		ho	7×2	5×2+1×3	27
75		hu	4×2+2×3	6×2	26
76		hã	2×1+2×2+3×3	6×2	27
77		hɛ̃	2×1+8×3	7×1+4×3	45
78		hĩ	6×3	4×1+7×2	26
79		hɔ̃	7×2	6×2	26
80		hũ	5×3	4×1+3×3	28
81		dʒa	4×2	6×2	20
82		dʒɛ	3×1+2×2+4×3	2×1+3×2	27
83		dʒe	5×3	4×1+2×3	25
84		dʒi	3×3	2×2	13
85		dʒɔ	3×2+2×3	8×2	28
86		dʒo	2×1+1×2	—	4
87		dʒu	2×2+1×3	1×2+1×3	12
88		ka	1×2+1×3	1×2	7
89		kɛ	2×2+2×3	4×2	18
90		ke	4×2+4×3	3×1+2×2+2×3	39
91		ki	2×3	1×1+1×2	9
92		kɔ	7×2	6×2	26
93		ko	1×2+3×3	2×1+2×2+2×3	23
94		ku	1×1+2×3	2×1	9
95		kpa	3×2	3×2	12
96		kpɛ	1×2+4×3	4×1+2×2	22
97		kpe	2×1+2×2	1×2	8
98		kpi	3×2+4×3	4×1+6×2	34
99		kpɔ	2×2+4×3	5×2	26
100		kpo	4×2	4×2	16
101		kpu	2×2+4×3	2×1+5×2	28
102		kpã	2×2+2×3	2×1+4×2+1×3	23
103		kpɛ̃	2×1+1×2+4×3	4×1+2×2	24
104		la	4×2	—	8
105		lɛ	3×2	—	6
106		le	3×2+2×3	4×2	20
107		li	2×3+1×**2**	2×1	10
108		lɔ	3×3	2×1+1×2	13
109		lo	4×3	1×1+2×2+1×3	20
110		lu	1×2+1×3	1×2	7
111		ma	3×3	2×1+2×3	17
112		mɛ	4×2	—	8

	sign	transliteration	components	connections	complexity
113		me	2×2+3×3	2×1+3×2	21
114		mi	2×3	—	6
115		mɔ	2×2+2×3	2×1	12
116		mo	2×1+1×2+4×3+2×**2**	4×1+2×2	28
117		mu	2×2+1×3	2×2	11
118		mɓa	1×1+6×2	6×2	25
119		mɓɛ	2×1+1×2+1×3	—	7
120		mɓe	2×1+2×2+1×3	2×2	13
121		mɓi	2×1+2×2+4×3	4×1+4×2	30
122		mɓɔ	2×1+3×2+1×3	4×2	19
123		mɓo	1×2+4×3+2×**2**	4×1+2×2	26
124		mɓu	2×1+5×3	4×1+3×2	27
125		mgba	2×1+3×2	3×2	14
126		mgbɛ	2×1+1×2+4×3	4×1+2×2	24
127		mgbe	2×1+3×2	1×2+1×3	13
128		mgbɔ	2×1+2×2+4×3	5×2	28
129		mgbo	2×1+4×2	4×2	18
130		na	3×2	2×2	10
131		nɛ	2×3	2×3	12
132		ne	2×1+3×3	2×1+1×3	16
133		ni	1×2+4×3	2×1+2×2	20
134		nɔ	4×2+4×3	3×1+2×2+2×3	33
135		no	2×1+5×3	2×1+2×2+1×3	26
136		nu	2×2+2×3	2×1+4×2	20
137		nɗa	2×1+4×2+1×3	6×2	25
138		nɗɛ	4×2	—	8
139		nɗe	2×2+2×3	1×1+2×2+1×3	18
140		nɗi	8×3	4×1+2×2+1×3	35
141		nɗɔ	2×1+2×2+3×3	2×1+2×2+1×3	24
142		nɗo	6×3	3×1+3×2+1×3	30
143		nɗu	2×2+1×3	2×2	11
144		ɲa	4×2+2×3	6×2	26
145		ɲɛ	6×2+2×3	11×2+1×3	43
146		ɲi	1×2+2×3	1×1+1×2+1×3	14
147		ɲɔ	3×2+6×3	4×1+4×2	36
148		ɲʤa	3×2+2×3	11×2+1×3	37
149		ɲʤɛ	2×1+3×2+4×3	2×1+4×2	30
150		ɲʤe	3×1+3×2	2×2	13
151		ɲʤi	2×1+3×3	2×2	15

	sign	transliteration	components	connections	complexity
152		ɲdʒɔ	2×1+3×2+2×3	8×2	30
153		ɲdʒo	2×1+2×2	—	6
154		ɲdʒu	3×2+1×3	1×2+2×3	17
155		ŋa	2×1+3×3	2×1+1×3	16
156		ŋɛ	7×2	6×2	26
157		ŋɔ	2×1+2×3	—	8
158		ŋga	2×2+4×3	2×1+10×2	38
159		ŋgɛ	2×1+2×2+2×3	4×2	20
160		ŋge	2×2+3×3	2×1+2×2+1×3	22
161		ŋgi	1×1+2×3	1×1+1×2	10
162		ŋgɔ	4×1+2×2+3×3	4×2	25
163		ŋgo	2×2+3×3	2×1+3×2+2×3	27
164		ŋgu	2×1+2×2+2×3	2×1+2×2	18
165		pa	2×1+1×2+2×3	2×2	14
166		pɛ	4×2	3×2	14
167		pe	2×2+6×3	6×1+4×2	36
168		pi	5×2+2×**2**	7×2	28
169		pɔ	1×2+6×3	4×1+4×2	32
170		po	2×2+4×3	3×1+2×2	23
171		pu	4×2	4×3	20
172		ra	4×2+2×3	1×1	15
173		rɛ	3×2+2×3	1×1	13
174		re	3×2+4×3	1×1+4×2	27
175		ri	4×3+1×**2**	3×1	17
176		rɔ	5×3	3×1+1×2	20
177		ro	6×3	2×1+2×2+1×3	27
178		ru	1×2+3×3	1×1+1×2	14
179		sa	6×3	6×1+3×3	33
180		sɛ	1×2+3×3	2×2+1×3	18
181		se	3×2	—	6
182		si	3×2+4×3	4×1+4×2+1×3	33
183		sɔ	6×2	5×2	22
184		so	4×2+1×3	4×2	19
185		su	5×2	2×2	14
186		ʃa	2×1+6×3	6×1+3×3	35
187		ʃɛ	2×1+1×2+3×3	2×2+1×3	20
188		ʃe	4×2	—	8
189		ʃi	2×1+3×2+4×3	4×1+4×2+1×3	35
190		ʃɔ	2×1+6×2	5×2	24

	sign	transliteration	components	connections	complexity
191		ʃo	2×1+4×2+1×3	4×2	21
192		ʃu	2×1+5×2	2×2	16
193		ta	1×2+2×3	3×2	14
194		tɛ	3×2+4×3	4×1+5×2	32
195		te	3×2+2×3	4×2	20
196		ti	2×1+3×3	2×2	15
197		tɔ	4×2	3×2	14
198		to	3×1+1×3	—	6
199		tu	2×1+2×2+2×3	2×1+2×2	18
200		va	3×2+4×3	2×1+4×2	28
201		vɛ	6×2	6×2	24
202		ve	3×2+2×3	2×1+2×2+1×3	21
203		vi	6×3	1×1+6×2	31
204		vɔ	2×2+4×3	4×1+1×2+4×3	34
205		vo	3×2+2×3	3×2+2×3	24
206		vu	1×2+7×3	5×1+5×2	38
207		wa	3×3	2×2	13
208		wɛ	4×3	2×1+1×2+1×3	19
209		we	5×2+1×3	6×2	25
210		wi	1×2+6×3	1×1+5×2+1×3	34
211		wɔ	6×3	3×2	24
212		wo	6×2	5×2	22
213		wu	3×2+2×3	5×2	22
214		wã	1×2+4×3	4×2	22
215		ja	5×3	4×2	23
216		jɛ	3×2+4×3	2×1+4×2	28
217		je	3×1+2×2	1×2	9
218		ji	2×1+3×3	2×2	15
219		jɔ	2×1+4×3	4×1+1×2	20
220		jo	2×1+1×2	—	4
221		ju	2×1+2×2+1×3	1×2+1×3	14
222		za	8×3	8×1+2×2+4×3	48
223		zɛ	4×1+1×2+3×3	2×2+1×3	22
224		ze	5×2	—	10
225		zi	4×2+2×3	2×1+3×2	22
226		zɔ	7×2	5×2+1×3	27
227		zo	4×3	4×1+1×2	18
228		zu	6×2	3×2	18
229		ŋ	4×3	1×1+3×2	19

***** Bold numbers (**2**) correspond to the filled areas.

The distribution of complexities in all previously investigated scripts was uniform. Surprisingly, the hypothesis is rejected for the Vai script (cf. the following table).

Table 2
Distribution of complexities

C	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C
4	2	**9**	4	**14**	12	**19**	8	**24**	13	**29**	2	**34**	5	**39**	1	**44**	0
5	1	**10**	4	**15**	6	**20**	12	**25**	8	**30**	6	**35**	3	**40**	0	**45**	1
6	7	**11**	5	**16**	9	**21**	5	**26**	13	**31**	1	**36**	5	**41**	0	**46**	1
7	3	**12**	7	**17**	7	**22**	10	**27**	9	**32**	4	**37**	1	**42**	1	**47**	0
8	7	**13**	8	**18**	12	**23**	7	**28**	10	**33**	3	**38**	3	**43**	3	**48**	1

The uniformity hypothesis will be tested by the run test. Denote I the inventory size and R the range of complexities (for the Vai script we have $I = 229$ and $R = 44$). If the data are uniformly distributed, all expected frequency values are $E = \frac{I}{R+1}$. A run is a sequence of frequencies which are either all greater than E or all smaller than E. Hence we have $E = \frac{229}{44+1} = 5.09$ and 11 runs, namely [2,1, 7, 3, 7, 4,4,5, 7,8,12,6,9,7,12,8,12, 5, 10,7,13,8,13,9,10, 2, 6, 1,4,3,5,3,5,1,3,1,0,0,1,2,0,1,1,0,1]. Denote $n = R + 1$, n_1 the number of frequencies smaller than E and n_2 the number of frequencies greater than E (in this case $n = 45$, $n_1 = 26$, $n_2 = 19$). The number of runs is considered random (and, consequently, the distribution is considered uniform) if

$$z = \frac{|r - E(r)| - 0.5}{\sigma_r} < 1.96,$$

where r is the number of runs, $E(r) = 1 + \frac{2n_1n_2}{n}$ and $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$. We obtain $z = 3.55$, which means the Vai script is the first case where the uniform distribution does not yield a good fit.

Here, it is necessary to note the peculiarity of syllabic scripts in comparisons with alphabets. With the number of characters significantly higher, syllabaries usually contain some redundant signs as a result of additional filling in the gaps not occurring in original versions, cf. also syllable representations in the latest version of the Bamum script, another indigenous African invention (Schmitt 1963: Tab. 15; Mafundikwa 2004: 87–88). For the Vai script, the typical number of utilizable syllables hardly reaches a hundred (Singler 1996). That is, it would be interesting to check the uniformity hypothesis having sufficiently long native Vai texts to separate the core of the syllabary and its marginal part. This task together with character frequency will be addressed in future works.

In previous works (Buk, Mačutek and Rovenchak 2008, Mačutek 2008) the numbers of components and connections were also investigated (for Latin, Cyrillic, and Runic scripts). The Poisson distribution ($P_x = e^{-\lambda}\lambda^x / x!, \lambda > 0$) was applied as a model in both cases. The parameter λ is the mean of the distribution, which leads to a quite straightforward interpretation – the numbers of components and connections are controlled by the respective means, a character with 'too many' components or connections occurs with a low probability. More-

over, as the parameter is also the variance of the distribution, the higher the mean, the higher variability is expected. A relatively high number of Vai characters without a connection makes it necessary to modify the respective model, the result being the hyper-Poisson distribution ($P_x = a^x / {}_1F_1(1; b; a)\, b^{(x)}$, $a \geq 0$, $b > 0$, ${}_1F_1(1; b; a)$ is a hypergeometric function). We remind that the Poisson distribution is its special case for $b = 1$, cf. Wimmer and Altmann (1999). The Vai script with its 229 characters provides another corroboration of the models.

Table 3

Numbers of components and connections

	components	**connections**
0		17
1		9
2	7	30
3	22	24
4	41	33
5	49	33
6	48	34
7	35	16
8	17	10
9	8	10
10	2	5
11		3
12		4
13		0
14		1
	Poisson $\lambda = 3.50,\ \chi^2 = 4.39$ $P = 0.73,\ DF = 7$	Hyper-Poisson $a = 10.73,\ b = 7.50$ $\chi^2 = 18.86$ $P = 0.09,\ DF = 12$

Acknowledgement

J. Mačutek was supported by the research grant VEGA 1/3016/06.

References

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

Altmann, G. (2008). Towards a theory of script. In: Altmann, G., Fan, F. (eds.), *Analyses of Script. Properties of Characters and Writing Systems: 149-164.* Berlin: de Gruyter.

Bird, S. (1999). Strategies for Representing Tone in African Writing Systems. *Written Language and Literacy 2, 1–44.*

Buk, S., Mačutek, J., Rovenchak, A. (2008). Some properties of the Ukrainian writing system. *Glottometrics 16, 63-79.*

Census (2008). *2008 National Population and Housing Census. Preliminary results.* Monrovia: Liberia Institute of Statistics and Geo-Information Services.

Coulmas, F. (2004). *The Blackwell Encyclopedia of Writing Systems.* Blackwell Publishing.

Dalby, D. (1967). A survey of the indigenous scripts of Liberia and Sierra Leone: Vai, Mende, Loma, Kpelle and Bassa. *African Language Studies 8, 1–51*.

Everson, M., Nyei, M., Riley, Ch., Sherman, T. (2006). Proposal for addition of Vai characters to the UCS. http://www.dkuug.dk/jtc1/sc2/wg2/docs/n3081.pdf.

Galtier, G. (1987): Un exemple d'écriture traditionnelle mandingue: le "Masaba" des Bambara-Masasi du Mali. *Journal des Africanistes 57(1), 255–266.*

Gordon, R. G., Jr. (ed.) (2005). Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com.

Jensen, H. (1969). *Die Schrift in Vergangenheit und Gegenwart. 3., neubearb. und erw. Aufl*. Berlin: Deutscher Verl. der Wissenschaften.

Köhler, R. (2008). Quantitative analysis of writing systems: An introduction. In: Altmann, G., Fan, F. (eds.), *Analyses of Script. Properties of Characters and Writing Systems: 3-9.* Berlin: de Gruyter.

Mačutek, J. (2008). Runes: complexity and distinctivity. *Glottometrics 16, 1-16.*

Mafundikwa, S. (2004). *Afrikan Alphabets: The Story of Writing in Afrika*. New York: Mark Butty Publisher.

Massaquoi, M. (1911). The Vai people and their syllabic writing. *Journal of the Royal African Society 10(40), 459–466*.

Ofri-Scheps, D. (1991). Vai Phonemics. In: *On the Object of Ethnology: Apropos of Vai Culture in Liberia: 84–119.* Ph. D. Dissertation, University of Berne.

Priest, L. A. (2004). Vai Syllabary. http://scripts.sil.org/cms/scripts/render_download.php?site_id= nrsi&format=file&media_id=VaiUnicode&filename=Vai+Syllabary2004-12-01.pdf.

Schmitt, A. (1963). *Die Bamum-Schrift: 3 Bde.* Wiesbaden: Otto Harrassowitz.

Singler, J. V. (1983). [Review of] The Psychology of Literacy. By Sylvia Scribner and Michael Cole. *Language 59(4), 893–901*.

Singler, J. V. (1996). Scripts of West Africa. In: Daniels, P. T., Bright, W. (eds.), *The World's Writing Systems: 593–598*. Oxford: Oxford University Press.

Tuchscherer, K. (2005). History of writing in Africa. In: Apiah, K. A., Gates, H. L., Jr. (eds.), *Africana: The Encyclopedia of the African and African American Experience (second edition): 476–480.* New York: Oxford University Press.

Tuchscherer, K. (2007). Recording, communicating, and making visible: A history of writing and systems of graphic symbolism in Africa. In: *Inscribing Meaning. Writing and Graphic Systems in African Art: 37–53*. Smithsonian.

Tuchscherer, K., Hair, P. E. H. (2002). Cherokee and West Africa: Examining the origins of the Vai script. *History in Africa 29, 427-486.*

Tucker, S. L. (1999). *[Vai kpolo lɔ kowae gbu kpaha sɔ kpoloe fɔlanaã mɛ = Vai script primer 2]*. Monrovia: Libtralo.

Vydrine, V. (1999). *Manding–English Dictionary (Maninka, Bamana). Vol. 1.* St. Petersburg: Dimitry Bulanin Publishing House.

Welmers, W. E. (1976). *A Grammar of Vai*. Berkeley: University of California Press.

Wimmer, G., Altmann, G. (1999). *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm.

Glottometrics 18, 2009, 13-25

Diversifikation deutscher morphologischer Klassen in SMS

Janja Laufer, Graz
Emília Nemcová, Trnava

Abstract. The aim of this article is to examine some properties of the SMS language, namely the rank-frequency sequences of some classes, and to elucidate whether they abide by some general interlinguistically valid principles.

Keywords: SMS, rank-frequency sequences, word classes

Einführung

Wortarten betreffende Erscheinungen bilden üblicherweise paradigmatische Klassen, z.B. die Klassen einer grammatischen Kategorie, oder lexikalische Klassen, z.B. Verbtypen. Die Kriterien ihrer Etablierung sind entweder morphologisch, syntaktisch, semantisch, soziolektal u.ä. oder gemischt. Die Etablierung der Wortklassen ist jedoch immer vom Blickwinkel oder dem Zweck der Untersuchung abhängig. Bei praktischen Zwecken bemüht man sich die Entitäten so zu ordnen, dass für die ganze Klasse die gleiche grammatische Regel gilt. Bei theoretischen Zwecken etabliert man die Klassen so, dass eine oder mehr Eigenschaften der Elemente der Klasse einer Hypothese folgen. Es wird angenommen, dass Eigenschaften sprachlicher Entitäten (Individuen oder Klassen) von Mechanismen kontrolliert werden, die man mit Gesetzeshypothesen erfassen kann. Diese Sicht liefert ein unabhängiges Kriterium der Klassenbildung: eine Klasse ist dann adäquat etabliert, wenn eine ihrer Eigenschaften einem Gesetz folgt bzw. nur die Entitäten gehören zu der gegebenen Klasse, die dem Gesetz folgen.

Es ist zweckmäßiger von Gesetzes*hypothesen* zu sprechen, weil die Etablierung eines Gesetzes eine sehr langwierige Arbeit werden kann und die Existenz einer Theorie voraussetzt.

In diesem Beitrag werden wir uns mit der „Sprache der SMS“ beschäftigen. Kurzmitteilungen wurden schnell nach ihrer Etablierung von der Öffentlichkeit und einigen Sprachkritikern als Plattform des Sprachverfalls gesehen. In der Tat hat die in Kurznachrichten verwendete Sprache ihre Eigenarten und bildet genauso wie Idiolekt, Soziolekt, Argot, Slang u.ä. ein neues, angeblich sehr inhomogenes „-lekt“. Die Inhomogenität – soweit sie aufspürbar ist – beruht jedoch vor allen Dingen darauf, dass SMS kurze Äußerungen sind und von vielen Personen stammen. Ansonsten kann sie in jedem sprachlichen Phänomen versteckt sein, unabhängig davon, mit welchem Medium die Äußerung getätigt und überbracht wurde und ihre Aufdeckung würde eine unendliche Arbeit bedeuten.

In diesem Beitrag möchten wir nur feststellen, ob bestimmte kategorische Klassifikationen im Bereich der Wortklassen (z.B. Wortarten, Verbarten) nach einer a posteriori Rangierung den bekannten linguistischen „Rang-Gesetzen“ folgen. Falls ja, dann stellt SMS in dieser Hinsicht keine „abnormale“ Sprache dar; falls nein, dann kann man erst danach fragen, ob die Klassen adäquat aufgestellt wurden.

Korpus und Klassen

Die Untersuchung von SMS ist in ihren Anfängen und die einzelnen Forscher benutzen eigene Korpora kleineren oder größeren Umfangs (vgl. Androutsopoulos, Schmidt 2002; Arlt 2006; Döring 2002; Dürscheid 2002a; Höflich, Rössler 2001, Krause, Schwitters 2002, Schlobinskli u.a. 2001). Die Analyse in dem vorliegenden Beitrag beruht auf einem eigens erstellten Korpus. Wir haben uns auf den Bereich von Graz eingeschränkt, um mögliche dialektale Überlagerungen zu vermeiden. Es wurden 1296 SMS analysiert, davon 530 von weiblichen und 766 von männlichen Schreibern. Bei der Analyse wurden geschlechterspezifische Unterschiede jedoch nicht berücksichtigt. Die Befragten waren Jugendliche zwischen 13 und 17 Jahren, da die Frequenz der SMS-Verwendung bei Jugendlichen am größten ist. Das dieser Arbeit zugrunde liegende Korpus wurde mittels Fragebogen und USB-Kabel-Verbindung erhoben.

Um eine Rang-Frequenz-Analyse der Wortklassen und Unterklassen durchführen zu können, müssen alle derartigen im Text beobachteten Entitäten in relativ einheitliche Kategorien eingeteilt werden. Wie aber aus der germanistischen Literatur bekannt, gibt es keine Methode, mit deren Hilfe man es durchführen könnte. Dies hat mehrere Gründe: (1) Spracheinheiten haben eine vage Identität (vgl. Altmann 1996). In zahlreichen Fällen erfolgt die Identifizierung mit Hilfe von externen Kriterien, die oft den Charakter einer Entscheidung haben. So ist der Fall z.B. bei deutschen abtrennbaren Präfixen, bei zusammengesetzten Verben, die früher zusammen, heute getrennt geschrieben werden, usw. (2) Auch wenn es gelingt eine Entität zu identifizieren, ihre Zuordnung zu einer Klasse ist nicht immer kategorisch sondern fuzzy. Das heißt, die Zugehörigkeit einer Einheit zu einer Klasse ist Sache des Maßes. Das Maß lässt sich auf eine 0-1-Entscheidung reduzieren, was mit Informationsverlust verbunden ist. Klassenzugehörigkeit kann in Sprachen phonologisch, graphematisch, morphologisch oder syntaktisch markiert sein (Dt. *rennen, Rennen*, Eng. *the hand, to hand,* Dt. *schnell* als Adjektiv oder Adverb), dies ist jedoch nicht immer der Fall. (3) Die Kriterien sind keine dateninhärente Größen, sondern reine Konventionen, die von uns aufgestellt werden. Sie lassen sich nicht einheitlich an alle Sprachen anwenden, sondern haben lokalen Charakter. Sie werden zu einem bestimmten Zweck formuliert.

Das Ranghäufigkeitskriterium, das hier benutzt wird, ist kein Identifikations- oder Zuordnungskriterium, sondern nur ein Test der Zuordnungsadäquatheit. Es kann nicht auf einzelne Wörter angewandt werden, sondern nur an die Zusammensetzung der Klasse.

Unsere Analyse stützt sich auf die Einteilung der Wörter in fünf Wortarten und acht Wortklassen. Innerhalb der einzelnen Klassen kann man weitere Unterteilungen vornehmen. Diese Wortklassen sind: das Verb, das Substantiv, das Adjektiv, der Artikel, das Numeral[1], das Adverb, die Partikel und das Pronomen. Die Interjektionen werden zunächst den Partikeln zugeordnet.

Die Wortarten und Wortklassen werden weitläufig durch zwei Kriterien abgegrenzt, durch das syntaktische Kriterium und das Kriterium der Bedeutungstypen. Das syntaktische Kriterium bezieht sich auf die Stellung der einzelnen Wörter im Satz, das zweite Kriterium bemüht die vier grundlegenden Typen der Bedeutung: die kategorematische, die deiktische, die kategorielle und die synkategorematische. Beide Kriterien können nicht in jedem Fall Wörter eindeutig einer Klasse zuordnen, jedoch geben

[1] Die Problematik der Etablierung des Numerals als Wortklasse soll an dieser Stelle nicht erörtert werden, die dieser Analyse zugrunde liegende Wortarten/-klassen-Differenzierung bezieht sich auf Hentschel/ Weydt (2003).

sie, im Vergleich mit der Einteilung nach morphologischen Kriterien, die meist auf der Grunddifferenzierung von flektierenden und nicht flektierenden Wortarten basiert, eine differenziertere Möglichkeit Wortarten zu beschreiben.

Regularität

Wenn eine linguistische Klasse adäquat zusammengesetzt wurde, erwartet man, dass die Ranghäufigkeitssequenz der geordneten Häufigkeiten eine Regularität aufweist. Diese Regularität wäre gleichzeitig auch ein Zeichen dafür, dass die Klasse hinreichend homogen ist. Die Ranghäufigkeitssequenz pflegt man üblicherweise mit der Zipfschen (zeta) Sequenz oder Verteilung, oder mit einer Verallgemeinerung wie z.B. Zipf-Mandelbrotsches Gesetz, Lerch-Verteilung usw. zu erfassen. Neuere Forschung hat gezeigt (vgl. Altmann 1992; Popescu, Altmann, Köhler 2008), dass diese homogenisierten Klassen eher mit einer Überlagerung von Verteilungen oder Sequenzen besser zu erfassen sind. Dem Vorschlag folgend lässt sich die Ranghäufigkeitssequenz mit einer Überlagerung von Exponentialfunktionen der Form

$$(1) \qquad f(r) = 1 + A_1\exp(-r/a_1) + A_2\exp(-r/a_2) + \ldots$$

gut erfassen, wobei die einzelnen Komponenten die jeweilige Schicht erfassen. Benutzt man mehr Komponenten als es tatsächliche Schichten gibt, dann werden die Exponenten gleich und man kann die Komponente auslassen. Da man $\exp(-r/a) = 1/\exp(r/a) = 1/\exp(1/a)^r$ schreiben kann und 1/exp(1/a) = q setzen kann, entspricht die obige Formel (1)

$$(2) \qquad f(r) = 1 + A_1q_1^{\,r} + A_2q_2^{\,r} + \ldots$$

d.h. einer Überlagerung von geometrischen Folgen. Eine einfache geometrische Folge zur Erfassung der rangierten Phonemhäufigkeit hat bereits B. Sigurd (1968) benutzt, jedoch ohne die Konstante 1, die konventionell als Asymptote eingesetzt wird, weil es keine kleinere Häufigkeit als 1 gibt.

Betrachten wir zuerst die Wortklassen, deren Häufigkeiten in rangierter Form in Tabelle 1 dargestellt sind.

Tabelle 1
Wortklassenhäufigkeiten in SMS

Rang r	Klasse	Häufigkeit f(r)	Theoretisch f(r)
1	Verben	2815	3122.11
2	Partikeln	2550	2463.08
3	Pronomina	2416	1943.20
4	Substantive	1606	1533.09
5	Adverbien	1459	1209.58
6	Adjektive	767	954.39
7	Artikel	541	753.07
8	Numeralia	175	594.27

Die theoretischen Häufigkeiten ergeben sich aus der Formel

$$f(r) = 1 + 3956.5603(0.7888)^r$$

oder alternativ und identisch aus

$$f(r) = 1 + 3956.5603\exp(-0.2372r)$$

und der Determinationskoeffizient ergibt den hohen Wert von $R^2 = 0.91$. Auch wenn es in einigen Klassen – besonders den äußeren – Diskrepanzen gibt, kann man diese Klassifikation vorläufig annehmen. Die Diskrepanzen lassen sich eventuell dadurch lösen, dass man die Definitionen einzelner Klassen revidiert, aber dazu muss man unterschiedliche deutsche Texte verwenden. Es besteht auch die Möglichkeit, dass die SMS-Sprache gerade an diesen Punkten von der Standardsprache abweicht. Dies ließe sich durch Vergleich feststellen.

Der einfachste Test für Klassenhomogenität im Sprachgefüge ist die Zugabe der zweiten Komponente in (1) oder (2). Passen wir die Zweikomponentenvariante an, so erhalten wir

$$f(r) = 1 + 1962.8567\exp(-0.2372r) + 1993.7037\exp(-0.2372r)$$

mit demselben Determinationskoeffizienten und gleichen Werten von $f(r)$. Wenn die Exponenten beider Komponenten gleich sind – wie in diesem Fall $b = d = 0.2372$ – dann gibt es keine Inhomogenität. Wie man leicht feststellen kann, ergibt sich die Amplitude der einfachen Formel genau aus der Summe der beiden multiplikativen Konstanten, d.h. 1962.8567 + 1993.7037 = 3956.5604 (nach Rundung). Dieser Homogenitätstest zeigt, dass die oben benutzten Klassen aufeinander abgestimmt sind. Sollten sich in der Entwicklung von SMS größere Diskrepanzen (z.B. in den Extremklassen) ergeben, dann würde es bedeuten, dass sich die SMS-Sprache von der Standardsprache entfernt – ohne dass man sie mit der Standardsprache vergleichen muss.

Was *morphologische Verbtypen* betrifft, so gibt es auf der ersten Ebene nur 3 Klassen, nämlich unregelmäßige, z.B. *haben, gehen,* schwache, z.B. *smsen, shoppen, schauen* und starke, z.B. *schlagen, lesen*, wie in Tabelle 2 dargestellt.

Tabelle 2
Morphologische Verbtypen

1	Unregelmäßige Verben	1362	1347.66
2	Schwache Verben	840	883.74
3	Starke Verben	613	579.64

Die Rangkurve ergibt $f(r) = 2054.4052\exp(-0.4224r)$ mit $R^2 = 0.989$, was keineswegs verwundert, da man nur drei Klassen hat, die man mit zwei Parametern erfassen muss. Starke Verben sind noch in Übermaß vorhanden, daher kann man auch hier von keiner Diskrepanz mit der Standardsprache sprechen.

Eine andere, *funktionale Verbklassifikation,* ergibt fünf Klassen, nämlich Vollverben, z.B. *gehen, fragen, kommen*, Kopulaverben, z.B. *sein, werden*, Hilfsverben, z.B. *sein, haben, tun,* Modalverben (*dürfen, können, mögen, müssen, sollen, wollen*) und modifizierende Verben, z.B. *lassen, haben, nicht brauchen,* wie in Tabelle 3 dargestellt.

Tabelle 3
Funktionale Verbklassifikation

Rang	Verbtyp	f(r)	f(r), 1K	f(r), 2K
1	Vollverben	1956	1940.96	1955.96
2	Kopulaverben	323	443.91	357.10
3	Hilfsverben	296	102.12	242.70
4	Modalverben	230	24.09	165.05
5	Modifizierende Verben	10	6.27	112.35

Die Einkomponentenanalyse (1K) ergibt zwar ein zufriedenstellendes Resultat, nämlich

$$f(r) = 1 + 8497.2013\exp(-1.4771r)$$

mit $R^2 = 0.96$ und einem signifikanten F-Test (P = 0.003), aber die Diskrepanzen in den mittleren Klassen (r = 2,3,4) zeigen, dass entweder die ganze Klassifikation inhomogen ist oder es müssen Klassen zugegeben oder eliminiert werden. Die Zweikomponentenanpassung ergibt

$$f(r) = 772.886676\exp(-0.3875r) + 161666341\exp(-11.6354r)$$

mit einem verbesserten Determinationskoeffizienten $R^2 = 0.99$, jedoch einem nicht signifikanten F-Test ($P = 0.11$), einer enormen Amplitude und ungleichen Exponenten, die signalisieren, dass hier eine falsche Klassifikation oder eine starke Inhomogenität vorliegt. Man kann daraus folgern, dass man diese Klassifikation in den Grammatiken der Standardsprache eventuell überdenken sollte. Auf der anderen Seite besteht die Möglichkeit, dass in diesem Punkt die SMS-Sprache von der Standardsprache abweicht oder dass die Zuordnung bestimmter Verben zu einer dieser Klassen nicht adäquat ist.

Betrachtet man eine spezielle Klasse, nämlich die *Modalverben*, wie in Tabelle 4 dargestellt, so sieht man eine gute Übereinstimmung mit der Theorie die

$$f(r) = 141.9663\exp(-0.4655r)$$

mit $R^2 = 0.975$ liefert.

Tabelle 4
Modalverben in SMS

1	Können	92	90.13
2	Müssen	53	56.95
3	Wollen	33	36.13
4	Sollen	31	23.05
5	Mögen	17	14.85
6	Dürfen	5	9.69

Nimmt man die Verben als eine selbständige Klasse und untersucht eine ihrer Eigenschaften, nämlich *Tempus*, dann erhält man die Daten wie in Tabelle 5 dargestellt.

Tabelle 5
Tempora in SMS

1	Präsens	1902	1900.87
2	Perfekt	240	256.20
3	Präteritum	89	35.28
4	Futur I	38	5.61
5	Plusquamperfekt	3	1.62
6	Futur II	0	1.08

Die Anpassung ist sehr gut, da man von einer bekannten Diversifikation der Tempora ausgeht. Daher konnte auch die nicht vorkommende Klasse Futur II in Betracht gezogen werden. Die resultierende Formel ist

$$f(r) = 14144.0294\exp(-2.0075r)$$

die einen $R^2 = 0.9985$ liefert. Die berechneten Daten sind in der letzten Spalte der Tabelle 5.

Die *Modi* liefern, weil es nur 3 Klassen gibt, eine ebenso gute Anpassung, nämlich

$$f(r) = 32432.1592\exp(-2.7349r) \text{ mit } R^2 = 0.9998.$$

Das Resultat ist in Tabelle 6 zu sehen.

Tabelle 6
Modi in SMS

1	Indikativ	2106	2105.92
2	Imperativ	135	137.61
3	Konjunktiv	30	9.87

Etwas mehr Klassen liefern *Pronomina* als Unterklasse der Wortarten. Beispiele zu einzelnen Klassen sind: Personalpronomina: *ich, du, wir*; Indefinitpronomina: *ein bisschen, nichts, viele*; Demonstrativa: *der, diese, die*, Possessiva: *mein, dein, sein, unser*; Interrogativa: *wer, was*; Reflexiva: *sich, mich, uns*; Relativa: *der, welcher, die, welche*. Die resultierende Anpassung ist in Tabelle 7 angegeben. Es kann vermutet werden, dass hier auch zwei Schichten existieren, bzw. eine Inhomogenität vorliegt, weil akzeptable Werte nur die Zweikomponentenanpassung liefert, nämlich

$$f(r) = 1 + 97746.3839\exp(-4.2184r) + 304.9856\exp(-0.2273r)$$

mit sehr unterschiedlichen Parametern und einem $R^2 = 0.9988$. Die Resultate sind in Tabelle 7 enthalten. Die Version mit einer Komponente ergibt zwar einen hohen R^2-Wert, aber die mittleren Klassen weichen zu stark ab.

Tabelle 7
Pronomina in SMS

1	Personalpronomina	1683	1683.00
2	Indefinitpronomina	216	215.78
3	Demonstrativpronomina	148	155.55
4	Possessivpronomina	122	123.89
5	Interrogativpronomina	114	98.90
6	Reflexivpronomina	107	79.00
7	Relativpronomina	26	63.14

Die *Adverbien* haben wir nach Hentschel/Weydt (2003) syntaktisch in vier Klassen aufgeteilt, nämlich in Modaladverbien, z.B. *gern, sehr, so*; Satzadverbien, z.B. *leider, trotzdem, schon*; deiktische Adverbien, z.B. *hier, dort, gestern* und relationale Adverbien, z.B. *oft, rückwärts*. Die syntaktische Adverbienklassifizierung basiert auf dem syntaktischen Verhalten der Wortklasse, wobei drei Kriterien berücksichtigt werden müssen: 1) die Möglichkeit, dass das Adverb Fokus der Negation ist; 2) die Erfragbarkeit des Adverbs und 3) die Möglichkeit, dass das Adverb in einem negierten Satz steht, ohne selbst Träger der Negation zu sein.

Die Klassifikation ist in Tabelle 8 und die Resultate der Zählung in Tabelle 8a aufgeführt.

Tabelle 8
Klassifikation der Adverbien

Adverbart	Verweis auf / Bezeichnung	Negation	Erfragbarkeit	Träger der Negation
Deiktische Adverbien	Ort oder Zeit in der außersprachlichen Wirklichkeit	✓	✓	×
Relationale Adverbien	einer Eigenschaft, die nur im Verhältnis mit anderen besteht	✓	×	×
Modaladverbien	mit der semantischen Gruppe identisch	✓	✓	✓
Satzadverbien	heterogene Gruppe, ohne gemeinsamen Bezug	×	×	×

Tabelle 8a
Adverbien in SMS syntaktisch betrachtet

1	Modaladverbien	4189	4173.93
2	Satzadverbien	680	830.94
3	Deiktische Adverbien	541	166.06
4	Relationale Adverbien	49	33.82

Obwohl die Funktion

$$f(r) = 20981.4595\exp(-1.6150r)$$

einen guten Determinationskoeffizienten $R^2 = 0.98$ liefert, zeigt der t-Test, dass der Parameter *a* nicht signifikant ist ($P = 0.10$), obwohl die gesamte Anpassung signifikant ist (F-Test: $P = 0.008$). Da von vier Werten zwei stark abweichen, ist es anzunehmen, dass diese Klassifikation nicht ganz adäquat ist und möglicherweise mehr Klassen etabliert werden müssen. Auf der anderen Seite besteht die Möglichkeit, dass sich SMS gerade in diesem Punkt von der Standardsprache unterscheiden. Es müssten Analysen von anderen Texten durchgeführt werden, um eine Vergleichsbasis zu schaffen.

Ein anderes Bild liefert die semantische Klassifikation der Adverbien in 8 Klassen, nämlich Temporaladverbien: *jetzt, morgen, heute*; Modaladverbien: *gern, sehr, OK*; Interrogativadverbien; *wo, wann, wie, warum*; Lokaladverbien; *hier, dort, da, daheim*; Konsekutivadverbien: *sonst, andernfalls*; Konzessivadverbien: *trotzdem, außerdem*; Kausaladverbien: *deshalb, darum, deswegen* und Instrumentaladverbien: *dadurch, damit*. Das Resultat ist in Tabelle 9 dargestellt. Die semantische Klassifizierung ergibt sich durch die Art der durch Adverbien bezeichneten Umstände. Da hier mehr Klassen vorhanden sind, lässt sich die Anpassung leichter durchführen. Wir bekommen die Funktion

$$f(r) = 2620.3924\exp(-1.0928r)$$

mit $R^2 = 0.96$. Obwohl es in der Mitte (Klassen 2-4) größere Abweichungen gibt, was zu der Annahme führt, dass die Klassen nicht adäquat etabliert wurden, müssen wir auf Resultate aus anderen Texten warten, um den Hintergrund der Abweichung zu eruieren.

Tabelle 9
Adverbien in SMS semantisch betrachtet

1	Temporaladverbien	896	879.57
2	Modaladverbien	209	295.57
3	Interrogativadverbien	180	99.76
4	Lokaladverbien	123	34.11
5	Konsekutivadverbien	27	12.10
6	Konzessivadverbien	10	4.72
7	Kausaladverbien	9	2.25
8	Instrumentaladverbien	5	1.42

Die Klasse der *Partikeln* ist eine heterogene Klasse verschiedener „Reste“. Man kann sie eventuell als Synsemantika betrachten, zu denen die Pronomina nicht gehören. Hier haben wir Präpositionen (*von, für, mit, in, für, an*), Konjunktionen (*und, weil, dass, oder*), Abtönungspartikeln (*denn, ja, halt, eh*), Negationspartikeln (*nicht*), Antwortpartikeln (*ja, nein*), Konjunktionaladverbien (*deswegen, trotzdem*), Intensivpartikeln (*ziemlich, ganz, sehr*), Fokuspartikeln (*nur, bloß, allein*) und Modalwörter (*vielleicht, wahrscheinlich, sicher*).

Tabelle 10
Partikeln in SMS

1	Präpositionen	629	629.21
2	Konjunktionen	538	462.08
3	Abtönungspartikeln	249	328.46
4	Negationspartikeln	221	233.56
5	Antwortpartikeln	220	166.17
6	Konjunktionaladverbien	114	118.30
7	Intensivpartikeln	103	84.31
8	Fokuspartikeln	38	60.17
9	Modalwörter	21	43.02

Die resultierende Funktion zeigt, dass dies eine homogene Klasse ist. Alle Tests sind hoch signifikant, die einfache Kurve ergibt sich als

$$f(r) = 914.1168\exp(-0.3422r)$$

mit $R^2 = 0.95$. Die berechneten Werte sind in Tabelle 10 enthalten.

Das gesamte Resultat zeigt, dass die vorgeschlagene Theorie für rangierte Daten adäquat sein könnte. Die Annahme, dass es sich bei SMS um eine abweichende Sprache handelt, kann von dieser Sicht jedoch verworfen werden.

Diversifikation

Einem anderen Modell von Popescu folgend (vgl. Popescu, Mačutek, Altmann 2008; Fan, Popescu, Altmann 2008; Popescu, Altmann 2008a), der Ranghäufigkeitssequenzen als Diversifikationen einer bestimmten (Wort)Klasse betrachtet, lässt sich die „Güte" einer Klassifikation danach beurteilen, ob einige Eigenschaften dieser Sequenz zueinander in einem bestimmten Verhältnis stehen. Dieses Verhältnis wird als *c* bezeichnet. Die Untersuchung der SMS-Sprache kann eventuell ihre Abweichung vom Standarddeutsch zeigen. Für einige sprachliche Erscheinungen gibt es Intervalle um *c*, die aus 100 Texten in 20 Sprachen ermittelt wurden. Nach Popescus Annahme gilt die Beziehung

$$c = \frac{V + f_{\max} - f_{\min} + 1 - L}{h} = \frac{V + f(1) - f(V) + 1 - L}{h} \quad (3)$$

wo *V* die Zahl der Klassen ist (d.h. die höchste Rangzahl), *f*(1) ist die Häufigkeit der ersten (häufigsten) Klasse, *f*(*V*) die der letzten (seltensten) Klasse, *L* ist die Bogenlänge zwischen *f*(1) und *f*(*V*), die man als die Summe der Euklidischen Distanzen zwischen benachbarten Häufigkeiten berechnet, d.h. als

$$L = \sum_{r=1}^{V-1} [(f(r) - f(r+1))^2 + 1]^{1/2} \quad (4)$$

und *h* ist der bekannte *h*-Punkt (vgl. Popescu 2007) an der Sequenz, den man folgendermaßen berechnet:

$$(5) \qquad h = \begin{cases} r & \text{wenn es ein } r = f(r) \text{ gibt} \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{wenn es kein } r = f(r) \text{ gibt} \end{cases}$$

d.h., wenn es einen Wert gibt, bei dem der Rang *r* gleich der Häufigkeit *f*(*r*) ist, dann ist dies der *h*-Punkt. Gibt es keinen solchen Wert, dann empfiehlt es sich erst *f*(*V*)-1 von allen anderen Häufigkeiten zu subtrahieren und dann zwei solche benachbarten Häufigkeiten f_1 und f_2 und Ränge r_1 und r_2 zu nehmen, dass $f_1 > r_1$ and $f_2 < r_2$ (manchmal muss man nicht benachbarte Werte nehmen, was in Formel (5) bereits einkalkuliert wurde). Die Berechnung dieser Größen wird am Beispiel der Adverbien (syntaktisch) illustriert. In Tabelle 8 hatten wir

1	2	3	4
4189	680	541	49

Die Bogenlänge ergibt sich als

$$L = [(4189 - 680)^2 + 1]^{1/2} + [(680 - 541)^2 + 1]^{1/2} + [(541 - 49)^2 + 1]^{1/2} = 4140.$$

Bei der Berechnung von *h* sehen wir, dass die erste Bedingung nicht erfüllt ist, daher müssen wir den zweiten Teil von (5) berechnen. Erst subtrahieren wir (49 – 1) = 48 von allen Häufigkeiten und bekommen

1	2	3	4
4141	632	493	1

Hier sieht man, dass *f*(3) > 3 aber *f*(4) < 4, daher setzen wir diese Werte in (5) ein und bekommen

$$h = [493(4)\text{-}1(3)]/[4 - 3 + 493 - 1] = 3.9939$$

was man auf 4 runden kann. Setzt man nun diese Werte in (3) ein, so erhält man

$$c = \frac{4 + 4189 - 49 + 1 - 4140}{4} = 1.25.$$

Vergleichen wir diesen Wert mit der Tabelle von Popescu, Altmann (2008a), wie in Tabelle 11 gezeigt, so sehen wir, dass (1) die syntaktisch bestimmten Adverbien in deutschen SMS eindeutig zur Klasse von Synsemantika wie Präpositionen, Postpositionen, Konjunktionen u.a. gehören, was man in der sechsten Zeile und der letzten Spalte der Tabelle sieht. Unser *c* = 1.25 liegt ausschließlich in dem dort angegebenen Intervall. (2) Da dieses System sprachunabhängig gelten sollte, wird es durch die syntaktisch bestimmten Adverbien in SMS bekräftigt.

Tabelle 11
Popescus Koeffizient *c* für einige Sprachklassen
(Popescu, Altmann 2008a)

Kategorie	$\overline{c}$	s_c	**Intervall für c**	**Intervall für** $\overline{c}$
1. Laute, Phoneme, Buchstaben	1.05	0.02	<1.00, 1.10>	<1.04, 1.06>
2. Wortklassen (parts of speech)	1.10	0.02	<0.98, 1.29>	<1.08, 1.18>
3. Rhythmische Muster	1.14	0.11	<0.92, 1.36>	<1.10, 1.18>
4. Paradigmatische Klassen	1.15	0.05	<1.05, 1.26>	<1.10, 1.20>
5. Farbklassen	1.18	0.07	<1.05, 1.31>	<1.14, 1.22>
6. Präpositionen, Postpositionen, Konjunktionen	1.24	0.11	<1.03, 1.46>	<1.17, 1.32>
7. Kasusdiversifikation	1.33	-	-	-
8. Allomorphe des Plurals	1.37	0.21	<0.97, 1.77>	<1.31, 1.43>
9. Affixe (Bedeutungsdiversifikation)	1.39	0.16	<1.06, 1.71>	<1.33, 1,41>
10. Wörter (Bedeutungsdiversifikation)	1.47	0.21	<1.06, 1.88>	<1.44, 1.50>

Berechnen wir die *c*-Werte für die anderen oben behandelten Klassen, dann bekommen wir Resultate wie in Tabelle 12 angegeben.

Tabelle 12
Der *c*-Wert von Wortklassen in SMS

Klasse	V	L	h	f(1)	f(V)	c
Partikeln	9	608.53	8.56	629	21	1.11
Wortklassen	8	2640.01	8.00	2815	175	1.12
Pronomina	7	1657.17	6.93	1683	26	1.13
Modalverben	6	87.35	5.62	92	5	1.18
Funktionale Verbklassen	5	1946.03	4.98	1956	10	1.20
Tempora ohne 0	5	1899.03	4.85	1902	3	1.22
Adverbien syntaktisch	4	4140.00	3.99	4189	49	1.25
Adverbien semantisch	8	891.60	6.60	896	5	1.27
Modi	3	2076.01	2.98	2106	30	1.34
Morphologische Verbtypen	3	749.00	2.99	1362	613	1.34
Tempora mit 0	6	1902.19	4.97	1902	0	1.37

Unsere Resultate in Tabelle 12 vergleichen wir mit denen in Tabelle 11, wobei wir nur die Intervalle für den Wert *c* (nicht für seinen Mittelwert) in Betracht ziehen. Übereinstimmende Resultate bekommen wir in folgenden Fällen:

Wortarten in SMS liefern *c* = 1.12, das Konfidenzintervall ist <1.06, 1.15>. Die Unterklassifikationen einzelner Wortarten liefern:

Klasse	*c*	*Konfidenzintervall (zwischensprachlich)*
Verben, funktional	1.20	<1.05, 1.26>
Modalverben	1.18	<1.05, 1.26>
Pronomina	1.13	<1.05, 1.26>
Adverbien, syntaktisch	1.25	<1.05, 1.26>
Partikeln	1.11	<1.03, 1.46>

In zwei Fällen ergaben sich Differenzen, nämlich bei

Verben, morphologisch	1.34	<1.05, 1.26>
Adverbien, semantisch	1.27	<1.05, 1.26>

die aus dem Intervall fallen und entweder eine falsche Klassifikation oder einen Kontrast zwischen SMS und der Standardsprache signalisieren.

In zwei Fällen, nämlich bei grammatischen Kategorien Modus (c = 1.34) und Tempus (c = 1.37) ist kein Vergleich möglich, weil Popescu, Altmann (2008a) sie interlinguistisch nicht untersucht haben. Unsere Werte liefern jedoch zumindest das erste Bild über Diversifikation grammatischer Kategorien, die wahrscheinlich im Bereich der Kasusdiversifikation (Zeile 7) liegen wird. Weitere Analysen sind jedoch nötig, um dieses Phänomen auch zwischensprachlich zu erfassen.

Literatur

Altmann, G. (1992): Das Problem der Datenhomogenität. *Glottometrika 13, 287-298.*

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

Anonym (2003a): Britische Schülerin schreibt Aufsatz im SMS-Stil. In: *Der Standard vom 03.03.2003.*

Anonym (2003b): Die SMS lebt. In: *Der Standard vom 30.07.2003.*

Anonym (2007a): Briten sind SMS-Könige. In: *Der Standard vom 01.10.2007.*

Anonym (2007b): SMS-Sprache: Iren fürchten Verblödung der Jugend. In: *Der Standard vom 02.05.2007.*

Androutsopoulos, J., Schmidt, G. (2002): SMS-Kommunikation: Ethnographische Gattungsanalyse am Beispiel einer Kleingruppe. Zeitschrift für Angewandte Linguisitk *36, 49-80.*
URL: www.ids-mannheim.de/prag/sprachvariation/tp/tp7/SMS-Kommunikation.pdf. [Stand: 2007-05-10]

Androutsopoulos, J., Schmidt, G. (2004). löbbe döch. Beziehungskommunikation mit SMS. In: *Gesprächsforschung. Online-Zeitschrift zur verbalen Interaktion 5, 50-71*. URL: http://www.gespraechsforschung-ozs.de/heft2004/ga-schmidt.pdf [Stand: 2007-6-21]

Arlt, I. (2006): Zur Wortlängenverteilung in SMS-Texten. *Göttinger Beiträge zur Sprachwissenschaft 13*, 9-21.

Czarnota, T. (2003): SMS-Kommunikation – sprachliche und kommunikative Aspekte, Wien, Univ., Dipl.-Arb.

Döring, N. (2002): „Kurzm. wird gesendet" – Abkürzungen und Akronyme in der SMS-Kommunikation. *Muttersprache 112(2), 97-114*. URL: http://www.nicola-doering.de/publications/sms-kurzformen-doering-2002.pdf [Stand: 2007-11-24]

Dürscheid, C. (2002a): E-Mail und SMS – ein Vergleich. In: Dürscheid, C.v., Ziegler, A. (Hrsg.), *Kommunikationsform E-Mail: 93-114.* Tübingen: Stauffenburg.

Dürscheid, C. (2002b): SMS-Schreiben als Gegenstand der Sprachreflexion. (= Networx 28). Online im Internet: URL: http://www.mediensprache.net/networx/networx-28.pdf [Stand: 2007-8-28].

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 69-81.*

Hentschel, E., Weydt, H. (2003): Handbuch der deutschen Grammatik. 3., völlig neu bearb. Aufl. Berlin / New York: de Gruyter.

Höflich, J.R. (2003): Vermittlungskulturen im Wandel: Brief – E-Mail – SMS. In: Gebhardt, J., Höflich, J.R. (Hrsg.): *Vermittlungskulturen im Wandel. Brief – E-Mail – SMS: 39-61*. Frankfurt am Main [u.a.]: Lang.

Höflich, J.R., Gebhardt, J., Steuber, S. (2003): SMS im Medienalltag Jugendlicher. Ergebnisse einer qualitativen Studie. In: Höflich, J.R., Gebhardt, J. (Hrsg.), *Vermittlungskulturen im Wandel. Brief, E-Mail, SMS: 265-289*. Frankfurt am Main [u.a.]: Peter Lang.

Höflich, J.R., Rössler, P. (2001): Mobile schriftliche Kommunikation oder: E-Mail für das Handy. Die Bedeutung elektronischer Kurznachrichten (Short Message Service) am Beispiel jugendlicher Handynutzer. *Medien & Kommunikationswissenschaft 49(4), 437-461*. URL: http://visor.unibe.ch/ws04/medienthemen/docs/Hoeflich_sms.pdf [Stand: 2007-06-29]

Krause, M., Schwitters, D. (2002): SMS-Kommunikation – Inhaltsanalyse eines kommunikativen Phänomens. (= Networx 27). URL: http://www.mediensprache.net/networx/networx-27.pdf. [Stand:2007-06-11]

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 555-565*. Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 97-111.*

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

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

Schlobinski, P. [u.a.] (2001). Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten in der SMS-Kommunikation. (= Networx 22) Online im Internet: URL: http://websprache.net/networx/docs/networx-22.pdf. [Stand: 2007-05-10]

Schlobinski, P. (2003). SMS-Texte – Alarmsignale für die Standardsprache? URL: http://www.mediensprache.net/de/essays/2/ [Stand: 2007-05-08]

Schlobinski, P., Watanabe, M. (2003). SMS-Kommunikation – Deutsch/Japanisch Kontrastiv. Eine explorative Studie (= Networx 31). URL: http://www.mediensprache.net/networx/networx-31.pdf. [Stand: 2007-05-15]

Schmidt, M. (2001). *SMS easy. Die besten Emoticons, Sprüche und Kürzel*. München: Heyne (= Heyne-Bücher, 48).

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

Glottometrics 18, 2009, 26-31

Diversifikation des Phonems /r/ im Deutschen

Karl-Heinz Best, Göttingen

Abstract. In this paper the 1-displaced Thomas distribution has been fitted to the ranked distribution of /r/-variants in German. It brings a further corroboration of the hypothesis that diversification processes abide by laws.

Keywords: Diversification, German, phonemics

0. Das Diversifikationsgesetz

Ein wesentliches Forschungsfeld der Quantitativen Linguistik besteht darin, die Gesetzmäßigkeiten zu untersuchen, denen sprachliche Phänomene unterliegen. Die Notwendigkeit, mathematische Mittel zur Erforschung der Sprache einzusetzen, betonen u.a. Altmann (1978, 1985b) und neuerdings wieder Köhler (2005). Diese sind jedoch nicht Selbstzweck, sondern Mittel zur Verwirklichung des Hauptziels der Wissenschaft: „Das höchste Ziel jeder Wissenschaft ist die Erklärung der Phänomene (und damit auch die Möglichkeit zu ihrer Vorhersage)“ (Köhler 2005: 17). Erklärung und Prognose sind aber darauf angewiesen, dass man zeigen kann, dass ein beobachteter Zustand oder Prozess einem Gesetz folgt.

Die Quantitative Linguistik hat eine ganze Reihe von Gesetzeshypothesen entwickelt und fast alle auch in mehr oder weniger umfangreichen Erhebungen getestet und fast immer bestätigt gefunden; etliche dieser Gesetze sind in Best (2006) dargestellt und exemplarisch überprüft. Eine große Rolle spielt in diesem Zusammenhang das sog. Diversifikationsgesetz, das Altmann (1985a) erstmals und in Altmann (1991) erneut vorgestellt hat. Es besagt, dass Diversifikationsprozesse sprachlicher Entitäten sich als gesetzmäßig erweisen. Um ein Beispiel aus dem Deutschen anzuführen: Der Plural deutscher Substantive weist eine ganze Reihe von Allomorphen auf. Untersucht man nun, wie häufig die einzelnen Allomorphe verwendet werden, und bringt sie in eine Rangfolge dergestalt, dass das häufigste Allomorph Rang 1, das zweithäufigste Rang 2 usw. einnimmt, so lässt sich zeigen, dass den so gefundenen Rangfolgen ein Gesetz zugrunde liegt. Für die Plural-Allomorphe in Briefen Kleists (Brüers & Heeren 2004) erwies sich die geometrische Verteilung als geeignetes Modell; bei Kurzerzählungen Schnurres konnte die negative hypergeometrische Verteilung mit Erfolg angewendet werden (Meuser, Schütte, & Stremme 2008). Dass hier verschiedene Modelle für das Diversifikationsgesetz infrage kommen, darf man Faktoren wie Zeit, Autor oder Textsorte zuschreiben. Die deutschen Plural-Allomorphe sind nur ein Anwendungsfeld von vielen; weitere finden sich in der angegebenen Literatur (Altmann 2005, Best 2006) und vor allem auch in Rothe (1991). Zusätzlich sei erwähnt, dass auch Eigennamen diesem Gesetz unterliegen (Best 2007).

1. Die /r/-Allophone des Deutschen als Diversifikationsphänomen

Sucht man einmal auf der phonetisch-phonologischen Ebene nach Diversifikationserscheinungen, so drängt sich im Fall des Deutschen das Phonem /r/ aufgrund seiner vielgestaltigen Erscheinungsformen auf. Es kann alveolar und uvular als Vibrant oder Frikativ gesprochen werden, erscheint aber in bestimmten Positionen auch vokalisiert in silbentragender Funktion

oder auch im Silbenrand. In wieder anderen Fällen wird es ausgelassen und zeitigt als Reflex eine Dehnung des Nachbarvokals. Solche Vielgestaltigkeit ist eine gute Voraussetzung dafür, die verschiedenen Erscheinungsformen in ihrer Häufigkeit zu erfassen, in eine Rangfolge zu bringen und dann auf ihre Gesetzmäßigkeit hin zu überprüfen. Die Hypothese, die hier zu prüfen ist, heißt also: Die verschiedenen /r/-Varianten treten, in eine Häufigkeitsrangfolge gebracht, gesetzmäßig auf. Entsprechende Daten verdanken wir Ulbrich (1972), der die Variation des Phonems /r/ anhand der Aussprache durch professionelle Sprecher, 25 Rundfunksprecher und 15 Schauspieler, untersucht hat. Die jeweilige /r/-Variante wurde von geschulten Personen bestimmt. Diese Bestimmungen bilden die Datenbasis der Tabellen in Ulbrichs Darstellung (Ulbrich 1972: 31, 37). Es handelt sich dabei um folgende /r/-Varianten (Ulbrich 1972: 129)[1]:

/r/ als Frikativlaut: „parkt“ /parkt/, phonetisch als [paʁkt];
/r/ vokalisch assimiliert (Ersetzung von /er/ oder /re/ durch [ɐ]): „Banner“ /baner/, phonetisch als [banɐ];
/r/ vokalisch substituiert (Ersetzung von /r/ durch [e̯]): „bohrt“ /bo:rt/, phonetisch als [bo:e̯t];
/r/ elidiert: „Marsch“ /marʃ/, phonetisch als [ma:ʃ];
/r/ als Vibrationslaut: „Rand“ /rant/ oder /Rant/, phonetisch als [rant] oder [Rant].

Die Werte für „r indifferent“ werden nicht berücksichtigt; damit sind solche Realisationen gemeint, die beim Abhören durch die genannten Personen unterschiedlich bestimmt wurden (Ulbrich 1972: 94); davon betroffen sind nur 2.5 – 3% der /r/-Realisierungen.

Die folgende Darstellung beschränkt sich auf die Gesamtübersicht (Ulbrich 1972: 129); sie ist die einzige Übersicht, in der Ulbrich die absoluten Beobachtungswerte angibt. Alle anderen Tabellen enthalten nur Prozentwerte; sie haben auch den Nachteil, oft für einzelne Varianten gar keine Beobachtungen anzugeben.

2. Anpassung des Modells an Ulbrichs Daten

Bei dem Versuch, ein geeignetes Modell für Ulbrichs Daten zu den /r/-Varianten zu finden, erwies sich die 1-verschobene Thomas-Verteilung als besonders erfolgreich. Bei diesem Modell handelt es sich um eine doppelte Poisson-Verteilung (Wimmer & Altmann 1999: 631). Sie muss in 1-verschobener Form angepasst werden, da in der Rangfolge keine Klasse $x = 0$ vorkommt. Dieser Wahl liegt die einfache Annahme zugrunde, dass die Varianten Poisson-verteilt sind; jedoch ist der Einfluss verschiedener Faktoren dermaßen heterogen, dass man die Poisson-Verteilung auf einer anderen Ebene verallgemeinern muss. Auf diese Weise entsteht aus der 1-verschobenen Poisson-Verteilung mit der wahrscheinlichkeitserzeugenden Funktion (WEF) G(t) = t*exp(a(t-1)) durch die Feller-Verallgemeinerung die 1-verschobene Thomas-Verteilung mit der WEF $H(t) = t * \exp\{a[t * \exp(b(t-1)) - 1]\}$, aus der sich durch schrittweise Ableitung die Wahrscheinlichkeitsfunktion

$$P_x = \begin{cases} e^{-a}, & x=1 \\ e^{-a} \sum_{j=1}^{x-1} \frac{a^j}{j!} \frac{(bj)^{x-1-j}}{(x-1-j)!} e^{-bj}, & x=2,3,4... \end{cases}$$

[1] Die /r/-Notierungen orientieren sich an denen in Best (2008: 3-5).

herleiten lässt. Sie wird an die /r/-Varianten der Rundfunksprecher und der Schauspieler angepasst, zuerst getrennt und dann in einer Gesamtdatei. Die Anpassungen wurden mit dem *Altmann-Fitter* (1997) durchgeführt. Die Ergebnisse:

Tabelle 1
/r/-Realisationen der Rundfunksprecher

x	/r/-Variante	f_x	NP_x
1	Frikativlaute	2968	2981.44
2	Vokalisch assimiliert	1610	1612.29
3	Vokalisch substituiert	1192	1126.86
4	Elidiert	542	600.25
5	Vibrationslaute	526	517.17
a = 0.8301 b = 0.4285 FG = 2 C = 0.0014			

Legende zu den Tabellen:

x Variantenklasse des /r/-Phonems
f_x beobachtete Anzahl der Realisierungen der entsprechenden Variantenklasse
NP_x aufgrund der Thomas-Verteilung berechnete Anzahl der Realisierungen dieser Variantenklasse
a, b Parameter der Thomas-Verteilung
FG Freiheitsgrade
C Diskrepanzkoeffizient, der mit $C \leq 0.01$ eine gute Übereinstimmung zwischen dem Modell und den Beobachtungsdaten anzeigt.

Der Diskrepanzkoeffizient wird hier als das geeignete Testkriterium angesehen, da die Zahl der Beobachtungen in allen drei Zusammenstellungen recht groß ist. Tabelle 1 zeigt mit dem Testergebnis von C = 0.0014, dass die Thomas-Verteilung sich als Modell für die Verwendung der /r/-Allophone überzeugend bewährt, wie auch die folgende Graphik verdeutlicht:

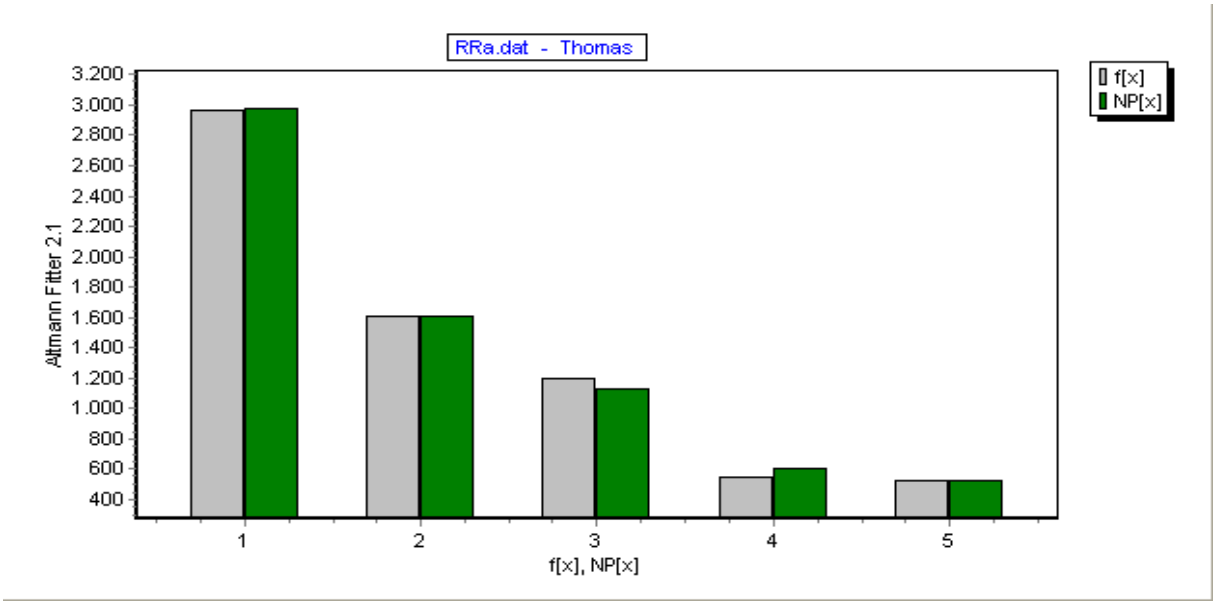


Abbildung 1. Anpassung der Thomas-Verteilung an die /r/-Realisationen der Rundfunksprecher

Für die /r/-Realisationen der Schauspieler ergab sich:

Tabelle 2
/r/-Realisationen der Schauspieler

x	/R/-Variante	f_x	NP_x
1	Frikativlaute	1161	1176.90
2	Vokalisch assimiliert	887	891.86
3	Vokalisch substituiert	686	650.17
4	Elidiert	372	376.64
5	Vibrationslaute	344	354.43
a = 1.0755 b = 0.3501 FG = 2 C = 0.0007			

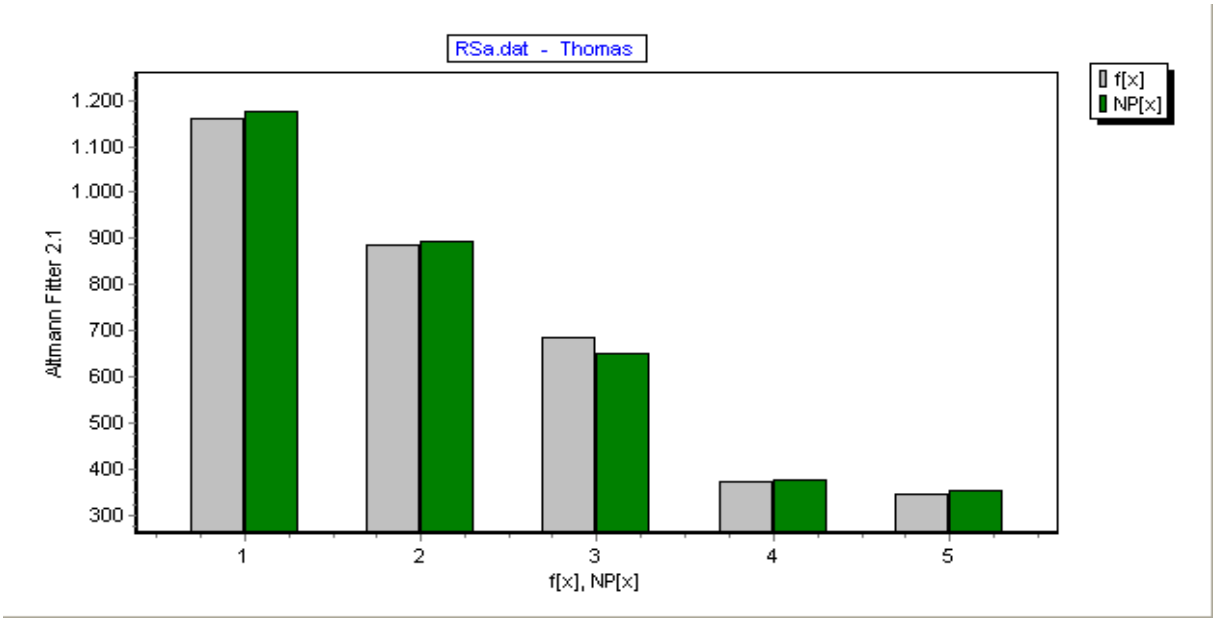


Abbildung 2. Anpassung der Thomas-Verteilung an die /r/-Realisationen der Schauspieler

Vereinigt man die Daten beider Sprechergruppen, so ergibt sich:

Tabelle 3
/r/-Realisationen der Rundfunksprecher und Schauspieler zusammen

x	/R/-Variante	f_x	NP_x
1	Frikativlaute	4129	4160.61
2	Vokalisch assimiliert	2497	2503.70
3	Vokalisch substituiert	1878	1775.88
4	Elidiert	914	975.27
5	Vibrationslaute	870	872.55
a = 0.9053 b = 0.4084 FG = 2 C = 0.0010			

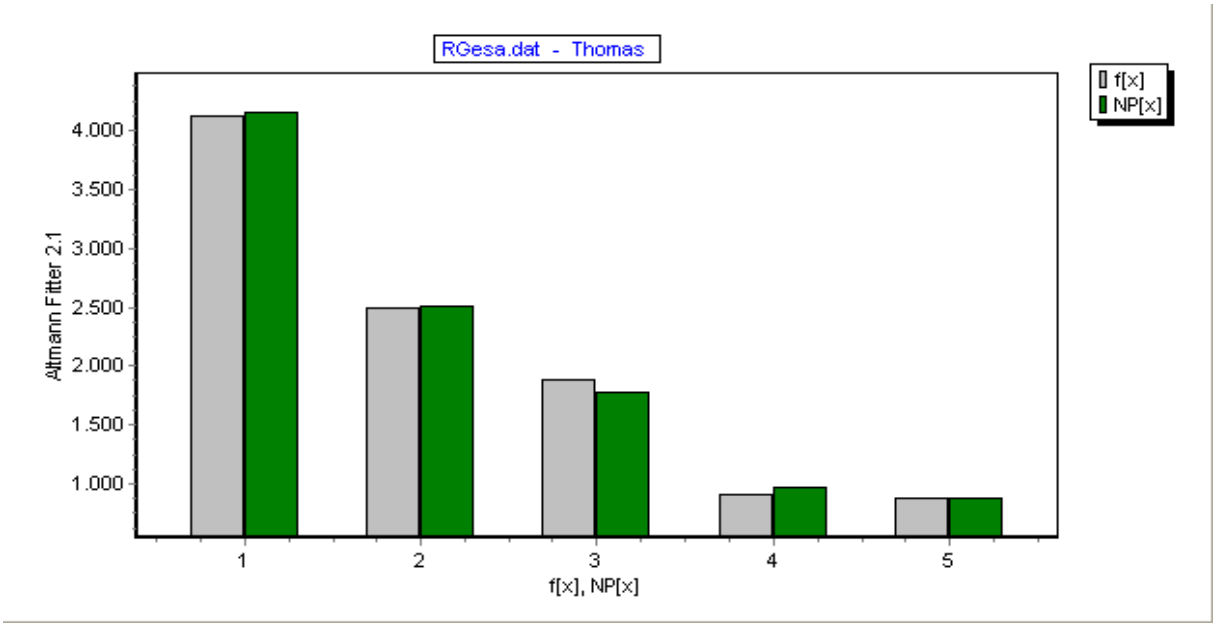


Abbildung 3. Anpassung der Tomas-Verteilung /r/-Realisationen der Rundfunksprecher und Schauspieler zusammen

Die Anpassung der Thomas-Verteilung an die Daten der /r/-Varianten ergab damit in allen drei Fällen sehr gute Ergebnisse. Die Hypothese, dass die Diversifikation des /r/ einem Sprachgesetz unterliegt, kann damit als bekräftigt angesehen werden.

3. Zusammenfassung

Die Untersuchung zeigt, dass auch auf phonetisch-phonologischer Ebene der Nachweis geführt werden kann, dass Diversifikationsprozesse sich gesetzmäßig verhalten. Derartigen Untersuchungen sind jedoch enge Grenzen gesetzt, da viele phonologischen Phänomene eine relativ geringe Variabilität aufweisen. Die meisten Phoneme erscheinen nur in der Form von ein oder zwei Varianten. In solchen Fällen kann die Diversifikation keinen Tests unterzogen werden, da die Verteilungsmodelle je nach der Zahl ihrer Parameter mindestens 3, oft aber mehr verschiedene Häufigkeitsklassen erfordern, um überhaupt getestet werden zu können. Umso bedeutsamer ist die Tatsache, dass die Gesetzeshypothese sich in einem Fall wie dem des Phonems /r/ mit seinen Varianten bewährt.

Die Wahl der Thomas-Verteilung als Modell muss nicht als endgültig betrachtet werden. Für die gleichen Daten konnte mit ebenfalls guten, nur unwesentlich schlechteren Ergebnissen auch die Polya-Verteilung verwendet werden, die allerdings mehr Parameter enthält und schon deshalb hier nicht bevorzugt wurde. Bei weiteren Untersuchungen auf phonetisch-phonologischer Ebene ist dieses Modell (und evt. auch noch weitere) jedoch ebenfalls in Betracht zu ziehen.

Literatur

Altmann, Gabriel (1978). Towards a theory of language. In: ders. (Hrsg.), *Glottometrika 1* (S. 1-25). Bochum: Brockmeyer.

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

Altmann, Gabriel. (1985b). Sprachtheorie und mathematische Modelle. *Christian-Albrechts-Universität Kiel, SAIS [= Seminar für Allgemeine und Indogermanische Sprachwissenschaft] Arbeitsberichte. H. 8, 1-13.*

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

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

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

Best, Karl-Heinz (2007). Diversifikation bei Eigennamen. In: Grzybek, Peter, & Köhler, Reinhard (Hrsg.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday* (S. 21-31). Berlin/ New York: Mouton de Gruyter.

Best, Karl-Heinz ([5]2008). *LinK: Linguistik in Kürze*. 5., durchgesehene Ausgabe. Skript. Lüdenscheid: RAM-Verlag.

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

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

Meuser, Katharina, Schütte, Jana, & Stremme, Sina (2008). Plural-Allomorphe in den Kurzgeschichten von W. Schnurre. *Glottometrics 17, 12-17.*

Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar*. Hagen: Margit Rottmann Medienverlag.

Ulbrich, Horst (1972). *Instrumentalphonetisch-auditive R-Untersuchungen im Deutschen.* Berlin: Akademie.

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

Software

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

Glottometrics 18, 2009, 32-39

Diversification of the case

Ioan-Iovitz Popescu, Bucharest
Emmerich Kelih, Graz
Karl-Heinz Best, Göttingen
Gabriel Altmann, Lüdenscheid

Abstract. The frequencies of individual grammatical cases in 4 languages are studied. Some indicators are used to compare their status with that of other language phenomena.

Key words: Diversification, case, German, Russian, Slovak, Slovenian

In a previous article (cf. Popescu, Altmann 2008) it has been shown that diversification of linguistic phenomena is a very regular process. The regularity is mirrored by the rank-frequency sequence of individual items of the diversified entity. It can be assumed that any kind of present diversification is a result of a mechanism which was present already in the initial genesis of language, that is, every diversification is an ontogenetic repetition of phylogenetic language evolution.

The start of diversification is caused by self-organization, a capability which is inherent not only in language. As soon as it is set in motion, Zipfian self-regulation intervenes and cares for some kind of equilibrium warranting successful communication. The only possibility to observe the regularity of the development of diversification consists in characterizing the rank-frequency sequence of the pertinent elements by some kind of indicator which signalizes a kind of constancy for every phenomenon. This indicator may be different for different phenomena in one language and different for the same phenomenon in different languages. The differences originate from different boundary conditions which exert specific influence on the phenomena.

Though up to now 12 different phenomena have been observed and the results turned out to be positive, this aspect of language is infinite. Individual investigations are necessary. In this contribution we shall restrict ourselves to the examination of the diversification of case in some languages where it is marked morphologically and has at least 4 forms. Case is considered in the sense of Latin grammar. Quite different classes of entities are present in e.g. Hungarian or Finnish where suffixes do not signalize only the Latin "case", or in Japanese where postpositions express many different relations just as prepositions do in English. Here we shall observe the frequency of individual cases (nominative, genitive, dative, accusative, vocative, locative, instrumental) in German, Slovenian, Russian and Slovak texts and after ranking the frequencies according to their amount we compute the following quantities:

Empirical arc length of the rank-frequency sequence as

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

where $f(r)$ is the frequency of the element at rank r, and $R = r_{max}$ is the highest rank. This is the sum of Euclidean distances between neighbouring elements; further the h-point defined as

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

For computing (2) 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 $f(R) > R$, then one must transform the whole ranked sequence in $f^*(r) = f(r) - f(R) + 1$. The h-point is a fixed point of the rank-frequency distribution having a number of properties that can be exploited in text analysis (cf. Popescu 2007; Popescu, Altmann 2006a,b, 2007, 2008; Popescu, Best, Altmann 2007; Mačutek, Popescu, Altmann 2007; Fan, Popescu, Altmann 2008).

Using the above indicators and the observed frequencies, one computes alternatively the indicators

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

where $f(1)$ is the greatest and $f(R)$ the smallest frequency and R is the highest rank, or

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

where $L_{max} = R - 1 + f(1) - f(R)$. Obviously, as can easily be seen, the indicators c and p are joined linearly by the relationship

$$(5) \qquad p = (ch - 2)/(h - 1)$$

Taking averages for a similar phenomenon in several languages the above quoted authors found the p, c and interval values as presented in Tables 1 and 2.

Table 1
The mean $\overline{p}$ and its interval (ranked by $\overline{p}$)
(from Popescu, Mačutek, Altmann 2009)

Category	$\overline{p}$	s_p	***p*-interval**	$\overline{p}$ **-interval**
1. Sounds, phonemes and letters	1.01	0.03	<0.96, 1.06>	<1.00, 1.02>
2. Word classes	1.02	0.09	<0.85, 1.20>	<0.96, 1.09>
3. Rhythmic patterns (Latin, Greek, German)	1.06	0.11	<0.84, 1.28>	<1.02, 1.10>
4. Pitches of 58 musical texts	1.09	0.10	<0.88, 1.29>	<1.06, 1.11>
5. Colour classes	1.09	0.07	<0.94,1.23>	<1.07, 1.10>
6. Allomorphs of German plural	1.11	0.22	<0.68, 1.53>	<1.04, 1.17>
7. Polish paradigmatic classes	1.11	0.05	<1.01, 1.21<	1.06, 1.16>
8. Auxiliaries	1.13	0.14	<0.88, 1.39>	<1.04, 1.22>

9. Word frequencies for German	1.17	0.10	<0.97, 1.36>	<1.15, 1.18>
10. Affixes (meaning diversification)	1.17	0.03	<1.12, 1.23>	<1.16, 1.18>
11. Words (meaning diversification)	1.19	0.19	0.81, 1.57>	<1.16, 1.22>
12. Word frequencies for 20 languages	1.22	0.13	<0.96, 1.48>	<1.20, 1.25>

Table 2
The mean $\overline{C}$ and its interval(ranked by $\overline{C}$)
(from Popescu, Mačutek, Altmann 2009)

Category	$\overline{C}$	s_c	***c*-interval**	$\overline{C}$ **-interval**
1. Sounds, phonemes, letters	1.05	0.03	<1.00, 1.10>	<1.04, 1.06>
2. Pitches of 58 musical texts	1.13	0.10	<0.94, 1.32>	<1.11,1.16>
3. Word classes (parts of speech)	1.14	0.08	<0.98, 1.30>	<1.08, 1.20>
4. Rhythmic patterns (Latin, Greek, German)	1.14	0.11	<0.92, 1.36>	<1.10, 1.18>
5. Polish paradigmatic classes	1.15	0.05	<1.05, 1.25>	<1.10, 1.20>
6. Colour classes	1.19	0.07	<1.05, 1.32>	<1.14, 1.23>
7. Word frequencies for German	1.22	0.09	<1.06, 1.39>	<1.21, 1.24>
8. Auxiliaries	1.24	0.11	<1.03, 1.45>	<1.17, 1.32>
9. Word frequencies for 20 languages	1.29	0.12	<1.06. 1.52>	<1.27, 1.31>
10. Allomorphs of plural	1.35	0.20	<0.97, 1.74>	<1.29, 1.41>
11. Affixes (Meaning diversification)	1.39	0.17	<1.06, 1.71>	<1.33, 1.45>
12. Words (Meaning diversification)	1.46	0.18	<1.10, 1.81>	<1.43, 1.48>

As can be seen, the indicators differ for different phenomena. The tables can be completed both by taking into account more languages and by considering further phenomena. Here the "case" will be examined.

In German, ten journalistic texts have been analyzed. For every text the cases were simply counted. A word appearing in a given case by grammatical agreement was not counted. In this way we obtained for the German text 01 the following sequence:

N D D N D D N N A D D D A N D N N N N N G D N D D N N A G N N A D D N N N D N
N N A D G D D N N N N D N D N A D N D D D N N D D D N A G D A N A N G D D N N
D N D G D A N N N D D D N N D N N D N A N D D D

yielding the following rank-frequency sequence

Rank	1	2	3	4
Case	N	D	A	G
Frequency	45	40	11	6

Since $f(R) > R$, namely 6 > 4, we perform the transformation $f^*(r) = f(r) - 6 + 1$ and obtain

Rank	1	2	3	4
Case	N	D	A	G
$f^*(r)$	40	35	6	1

from which the necessary quantities can easily be computed:

$L = [(45-35)^2 +1]^{1/2} + [(35-6)^2+1]^{1/2} + (6-1)^2+1]^{1/2} = 39.22$
$h = [6(4) – 1(3)] /[4 – 3 + 6 – 1] = 3.5$
$L_{max} = 4 – 1 + 45 – 6 = 42$

Inserting these values in (3) and (4) we obtain

$c = (4 + 40 – 1 + 1 – 39.22)/3.5 = 1.366$
$p = (42 – 39.22)/2.5 = 1.112$

Comparing these values with those in Tables 1 and 2 we see that *c* lays between place 10 and 11 and *p* between places 6 and 7, in the grammatical domain. In order to stabilize this value we compute the indicators for the other texts in German. The data and the indicators are given in Table 3.

Table 3
Rank-frequencies of German cases

Text	Data	*N*	*R*	*f*(1)	*f*(*R*)	*h*	*L*	L_{max}	*p*	*c*
01	45,40,11,6;	102	4	45	6	3.50	39.22	42	1.112	1.366
02	50,47,33,7;	137	4	50	7	3.89	43.22	46	0.962	1.229
03	48,44,27,10;	129	4	48	10	3.83	38.18	41	0.996	1.258
04	40,32,24,3;	99	4	40	3	3.86	37.15	40	0.997	1.256
05	30,28,27,4;	89	4	30	4	3.87	26.67	29	0.812	1.119
06	73,49,42,13;	177	4	73	13	3.90	60.11	63	0.997	1.254
07	58,32,26,4;	120	4	58	4	3.87	54.12	57	1.003	1.261
08	32,21,20,4;	77	4	32	4	3.82	28.49	31	0.890	1.181
09	51,39,34,7;	131	4	51	7	3.89	44.16	47	0.983	1.244
10	49,35,27,4;	115	4	49	4	3.88	45.12	48	1.000	1.258
11	78,53,28,2;	161	4	78	2	3.89	76.06	79	1.017	1.270
12	48,37,20,1;	106	4	47	1	3.85	46.11	49	1.014	1.270
13	48,45,35,3;	131	4	48	3	3.91	45.23	48	0.952	1.220
14	43,29,24,2;	98	4	43	2	3.87	41.16	44	0.990	1.251
15	32,26,23,3;	84	4	32	3	3.86	29.27	32	0.955	1.225
16	46,34,28,3;	111	4	46	3	3.88	43.14	46	0.993	1.253
17	56,46,38,2;	142	4	56	2	3.92	54.13	57	0.983	1.242
18	46,45,28,6;	125	4	46	6	3.87	40.47	43	0.882	1.171
19	43,39,34,4;	120	4	43	4	3.90	39.24	42	0.952	1.221
20	64,61,38,6;	169	4	64	6	3.91	58.20	61	0.962	1.228

The average $\overline{p} = 1.012$ and the standard deviation $s_p = 0.091$; the mean $\overline{c} = 1.231$ and the standard deviation $s_c = 0.067$.

In Slavic languages there are mostly 6 marked cases (seldom is there a seventh case, the vocative), i.e. a richer set of forms. In Table 4 we present Slovenian data, mostly private letters, and in Table 5 Slovak data, all taken from http://zlatyfond.sme.sk (October 1, 2008), one of the poems (Slk 09) is from the 19th century.

Table 4
Rank-frequencies of Slovenian cases

Text	Data	*N*	*R*	*f*(1)	*f*(*R*)	*h*	*L*	L_{max}	*p*	*c*
Sln 01	64,54,52,21,20,7;	218	6	64	7	5.64	57.75	62	0.916	1.108
Sln 02	89,65,49,39,23,13;	278	6	89	13	5.55	76.18	81	1.059	1.229
Sln 03	86,78,45,29,27,7;	272	6	86	7	5.76	79.37	84	0.973	1.151
Sln 04	81,64,42,26,17,15;	245	6	81	15	4.80	66.37	71	1.218	1.381
Sln 05	26,16,14,11,9,1;	77	6	26	1	5.45	25.75	30	0.955	1.147
Sln 06	82,77,52,31,27,16;	285	6	82	16	5.58	66.31	71	1.024	1.199
Sln 07	78,34,28,23,20,12;	195	6	78	12	5.44	66.42	71	1.032	1.210
Sln 08	43,34,28,15,14,5;	139	6	43	5	5.50	38.65	43	0.967	1.155
Sln 09	67,59,42,23,16,14;	221	6	67	14	4.75	53.43	58	1.219	1.383
Sln 10	133,97,91,48,28,16;	413	6	133	16	5.62	117.17	122	1.045	1.215

Here $\overline{p}$ = 1.041, s_p = 0.104, $\overline{c}$ = 1.218 and s_c = 0.094

Table 5
Rank-frequencies of Slovak cases

Text	Data	*N*	*R*	*f*(1)	*f*(*R*)	*h*	*L*	L_{max}	*p*	*c*
Slk 01	26,18,16,13,8,2;	83	6	26	2	5.29	24.6	29	1.016	1.202
Slk 02	14,13,5,4,2,2,2;	42	7	14	2	3.50	15.1	18	1.148	1.391
Slk 03	53,52,47,46,42,6;	246	6	53	6	5.86	48.1	52	0.811	1.014
Slk 04	36,31,30,25,10,5,4;	141	7	36	4	5.33	33.2	38	1.118	1.283
Slk 05	67,50,44,28,15,15,2;	221	7	67	2	6.57	66.2	71	0.858	1.032
Slk 06	39,36,22,12,10,4,2;	125	7	39	2	5.57	37.80	43	1.138	1.293
Slk 07	27,22,10,7,4,2,	72	6	27	2	4.50	25.70	30	1.229	1.400
Slk 08	28,20,10,9,7,3;	77	6	28	3	5.00	25.9	30	1.028	1.222
Slk 09	163,105,72,43,24,22,2;	431	7	163	2	6.71	161	167	0.993	1.143

Here $\overline{p}$ = 1.038 and s_p = 0.138, $\overline{c}$ = 1.220 and s_c = 0.139.

In Russian we took some novels of P.A. Čechov. Here 6 cases have been found. The results are presented in Table 6.

Table 6
Rank-frequency of Russian cases

Text	Data	*N*	*R*	*f*(1)	*f*(*R*)	*h*	*L*	L_{max}	*p*	*c*
Ru 01	134,74,43,24,19,15;	309	6	134	15	5.00	119.27	124	1.183	1.346
Ru 02	64,25,19,8,7,6;	129	6	64	6	3.92	58.97	63	1.380	1.538
Ru 03	36,31,28,27,6,3;	131	6	36	3	4.95	33.86	38	1.048	1.240
Ru 04	54,28,21,14,4,2;	123	6	54	2	4.82	52.45	57	1.191	1.359
Ru 05	35,15,15,12,7,4;	88	6	35	4	4.83	32.45	36	0.927	1.149
Ru 06	66,35,21,11,11,6;	150	6	66	6	5.17	61.20	65	0.911	1.122
Ru 07	36,28,28,22,17,15;	146	6	36	15	4.67	22.48	26	0.959	1.182
Ru 08	47,33,23,18,16,10;	147	6	47	10	5.29	37.50	42	1.049	1.229
Ru 09	83,31,21,15,12,12;	174	6	83	12	4.00	72.30	76	1.233	1.425
Ru 10	51,17,10,9,7,2;	96	6	51	2	5.17	49.84	54	0.998	1.191

The averages are $\overline{p}$ = 1.088 and $\overline{c}$ = 1.278, the standard deviations are s_p = 0.153, s_c = 0.134.

Resuming the above data we see that the “case”, a grammatical category diversifies in a very small interval. We obtained

	$\overline{p}$	$\overline{c}$
German	1.012	1.231
Slovenian	1.041	1.218
Slovak	1.038	1.220
Russian	1.088	1.278

We see that *p*, for the case, a marked grammatical category, lies between a purely semantic phenomenon (colour classes) and the markers of a category, namely plural which have $\overline{p}$ = 1.09 and 1.11 respectively, as can be seen in Table 1. The indicator *c* lies between the same phenomena (see Table 2) having $\overline{c}$ = 1.19 and 1.35 but shares this place with some other phenomena. Needless to say, many individual examinations will be necessary to elaborate on the whole hierarchy and to give it stability by taking into account many languages. Nevertheless, we are on the right track towards an attractor which is deeply concealed in texts and languages but exerts an influence on grammar and text formation.

Another remarkable phenomenon is the fact that the cases are used in different manner but the rank-order sequence maintains its above mentioned properties. For example, in Russian we have the following rank-orders for individual texts:

n g a l i d
n g a i d l
g a d n l i
n g a d l i
n g a i l d
n g a l d i
g n a i l d
n g l a i d
n g l a i d
n g a l d i

Out of 6! = 720 possible orders only 8 have been used here. That means, the use of cases in Russian is relatively stable. The stability could be characterized or tested using some nonparametric test, but such a procedure should be performed with a much greater amount of data. At the same time, one could state the historical stability of the given order, the changes in the dynamics of the given phenomenon and last but not least, it could be used for stylistic purposes: special texts can display different idiosyncrasies. This package of problems will be discussed in future.

References

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

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests for the *h*-point and related text characteristics. *Glottometrics 15, 42-52.*

Popescu, I.-I., Altmann, G. (2006a). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenschaft 13, 87-98.*

Popescu, I.-I., Altmann, G., (2006b). Some aspects of word frequencies. *Glottometrics 13, 2006, 23-46*.

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15, 42-52.*

Popescu, I.-I., Altmann, G. (2008a). Autosemantic compactness of text. In: Altmann, G., Zadorozhna, I., Matskulyak V. (eds.), *Problems in General, Germanic and Slavic Linguistics. Papers for 70th anniversary of Professor V. Levickij: 427-480*. Chernivtsi: Books – XXI.

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

Popescu, I.-I., Best, K.-H., Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics 14, 58-71.*

Popescu, I.-I., Mačutek, J., Altmann (2009). *Aspects of word frequencies*. (in preparation)

Texts

German

Text 01: „Nicht blind genug“ heißt Startverbot, ET (= Eichsfelder Tageblatt), 9.9.2008, S. 28, Sparte: „Sport“ .
Text 02: Es bleibt dabei: Mit links ist gut, ET, 9.9.08, S. 29, „Sport“.
Text 03: Serena Williams ist wieder am richtigen Platz. ET, 9.9.08, S. 29, „Sport“.
Text 04: Teuber: Hoffnungsträger und Vorbild zugleich. ET, 9.9.08, S. 28, „Sport“.
Text 05: Eiskalte Gieboldehäuser besiegen Pferdeberg. ET, 17.9.08, S. 27, „Sport“.
Text 06. Über Peking „kann man nur in Superlativen sprechen“. ET, 17.9.08, S. 28, „Sport“.
Text 07. Verletzter Czyz siegt für kranken Vater. ET, 17.9.08, S. 28, „Sport“.
Text 08: Werder enttäuscht Bremer Fans. ET, 17.9.08, S. 29, „Sport“,
Text 09. Schröder wartet zwei Stunden auf Gold. ET, 15.9.08, S. 20, „Sport“.
Text 10. Bötzel fehlt noch immer die Medaille. ET, 15.9.08, S. 20, „Sport“.

Deutsche Sagen. Hrsg. von den Brüdern Grimm. Berlin: Rütten & Loening 1984.
Text 11: Die drei Bergleute im Kuttenberg. S. 35f.
Text 12: Die Springwurzel. S. 41f.
Text 13: Die Schlangenjungfrau. S. 44f.

Text 14: Des kleinen Volks Hochzeitsfest. S. 58f.
Text 15: Zwerge leihen Brot. S. 61
Text 16: Das Bergmännlein beim Tanz. S. 65f.
Text 17: Der Wassermann. S. 73f.
Text 18: Die Elbjungfer und das Saalweiblein. S. 82f.
Text 19: Der Alraun. S. 120f. (1 lat. Zitat ausgelassen)
Text 20: Das Vogelnest. S. 124f. (1 lat. Wort ausgelassen)

Slovenian
Text 1-8: Cankar, Ivan (1898 – 1902): Private letters to Ana Lušinova. Ljubljana: DZS.
Text 9: Prežihov, Voranc (1940): Samorastniki. Chapter 1. (Novel). Ljubljana: Naša založba.
Text 10: Prežihov, Voranc (1940): Samorastniki. Chapter 2. Ljubljana. (Novel) Ljubljana: Naša založba.

Russian
All texts are from http://lib.ru/LITRA/CHEHOW/ (October 10, 2008)

Čechov, A.P.
Text 01: Chameleon. (1884).
Text 02: Ušla. (1983)
Text 03: Sovremennye molitvy. (1883).
Text 04: Sovet. (1883).
Text 05. Idillija. (1884)
Text 06: Na gvozde. (1883)
Text 07: Po-amerikanski. (1880)
Text 08: Radost'. (1883)
Text 09: Rjaženny e. (1883)
Text 10: Temnuju noč'ju. (1883)

Slovak
All texts are from http://zlatyfond.sme.sk (October 1, 2008)

Text 01: Ján Stacho, Apokryfy: Noc
Text 02: Rudolf Dilong: Nevolaj, nevolaj: Minieme sa.
Text 03: Ján Ondruš, Korenie: Chodec po povraze
Text 04: Ján Kovalik Ústiansky, Z pút k slobode: Bratom za Oceánom
Text 05: Anton Prídavok, Lámané drieky
Text 06: Jozef Gregor Tajovský, Zajac
Text 07: Pavol Ušák Oliva, Čierne kvietie: Hviezdy a smútok
Text 08: Lýdia Vadkerti-Gavorníková, Trvanie: Leto
Text 09: Janko Kráľ, Šahy. 1849

Glottometrics 18, 2009, 40-51

The End of Year Addresses of the Presidents of the Italian Republic (1948-2006): discoursal similarities and differences

Francesco Pauli[1]
Arjuna Tuzzi

Abstract. This work uses statistical and linguistic procedures to analyse a corpus of 57 traditional End of Year addresses by nine presidents of the Italian Republic. These addresses are compared and contrasted in order to identify the discoursal characteristic features and the similarities and differences among them. The proposed methodology is an attempt to link traditional qualitative methods with modern textual statistics.

Results show that the hypergeometric, χ^2 and bootstrap tests are effective in identifying distinctive words among presidents ('between' perspective) and among addresses delivered by the same president ('within' perspective) and that individual characteristic features and personal traits are more important than other factors in the End of Year Addresses topics.

Keywords: End of Year Addresses, lexicon, corpora, distinctive textual units, intertextual distance

1. Introduction

The President of the Italian Republic is the Head of State and represents the unity of the nation (Italian Constitution, par.87). In Italy, the Presidency of the Republic is the first Office of the State and is a strong Institution even from a symbolic point of view. In fact, in social imagery, the president represents the values and the genuine essence of the nation. He is the most authoritative – and often also most beloved – figure of the Italian leadership.

The traditional End of Year Address is a media event, a peculiar civil ritual, unique in its kind, because the president addresses directly the Italian citizens. Besides the obvious contents of good wishes and solemnity, the texts of the presidential addresses are a rich source of information on the last fifty years of Italian history. The comparison of the addresses can reveal the changing habits and morals of the country and of the personalities of the presidents. For this reason, an interdisciplinary research team of the University of Padova, composed of linguists, historians, politologists, sociologists and statisticians, has started a research project aimed at analyzing a corpus of presidential addresses from different disciplinary perspectives (Cortelazzo and Tuzzi 2007).

This article deals with the specific issue of identifying lexical features and discoursal similarities of the presidential end of year addresses by means of an explorative statistical analysis and Labbé and Labbé's intertexual distance. All the procedures are based on the analysis of contingency lexical tables containing word frequencies of corpus consisting of nine subcorpora. Text data were processed by means of the Taltac2 dedicated software

[1] Address correspondence to: F. Pauli, Dipartimento di Scienze Statistiche, via Battisti 241/243, I-35123 Padova. E-mail: fpauli@stat.unipd.it. Or to A. Tuzzi, Dipartimento di Sociologia, via Cesarotti 10/12, I-35123 Padova. E-mail: arjuna.tuzzi@unipd.it.

(www.taltac.it) and the statistical analysis was conducted by means of R (R Development Core Team 2007).

2. Corpus

Our corpus is composed of 57 addresses, delivered by nine presidents, from December 31st 1949, the date of the first address by Luigi Einaudi on the radio, to December 31st 2005, the date of the seventh and last address by Carlo Azeglio Ciampi on television (Table 1). The first address broadcast on television was the fifth by Giovanni Gronchi in 1959. Usually, each president give seven addresses since the duration of the presidential term is seven years. The two exceptions are Einaudi, whose first year (1948) address is not mentioned, and Segni, who delivered only two addresses because he resigned from his position for health problems after two years in office. Enrico De Nicola, Provisional Head of State and first President of the Republic (1946-48), is not considered in this analysis because the tradition of End of Year Addresses began with his successor, Luigi Einaudi.

Table 1
Number and length of the presidents' discourses.

President	period	no. of addr.	tokens	types (forms)	types (lemmas)
Luigi Einaudi	1948-55	6	1203	558	464
Giovanni Gronchi	1955-62	7	5829	1794	1388
Antonio Segni	1962-64	2	1795	781	669
Giuseppe Saragat	1964-71	7	8476	2238	1707
Giovanni Leone	1971-78	7	7379	2041	1516
Sandro Pertini	1978-85	7	15592	2743	1903
Francesco Cossiga	1985-92	7	13890	3001	2236
Oscar Luigi Scalfaro	1992-99	7	24675	4133	2811
Carlo Azeglio Ciampi	1999-06	7	12597	2919	2074
Corpus	*1948-06*	*57*	*91436*	*9786*	*6003*

The addresses are available in text and audio-visual format on the institutional web site of the presidential office (Quirinale, *www.quirinale.it*). Generally, the two versions differ slightly (Cortelazzo 1985), and so manual correction of the written texts downloaded from the site was needed in order to obtain the texts actually spoken by the presidents. In particular, the addresses delivered by Pertini are often different from the available written versions because of this president's habit of giving extemporaneous addresses. This is especially true for the 1984 address.

In this article, words are defined as sequences of letters taken from the alphabet and isolated by separators (blanks and punctuation marks). *N* is the corpus size in terms of total number of word-tokens. The frequency of a word-type is the number of corresponding word-tokens in the corpus; *V* is the vocabulary size in terms of different word-types. The length of the addresses shows a trend of increasing duration (Figure 1), the two extremes being Einaudi for his conciseness and Scalfaro for his loquacity. It is worth saying that in 1991 Cossiga, rather than the scheduled address, delivered a very short message introductory to his resignation from office a few months before the official end of office in 1992.

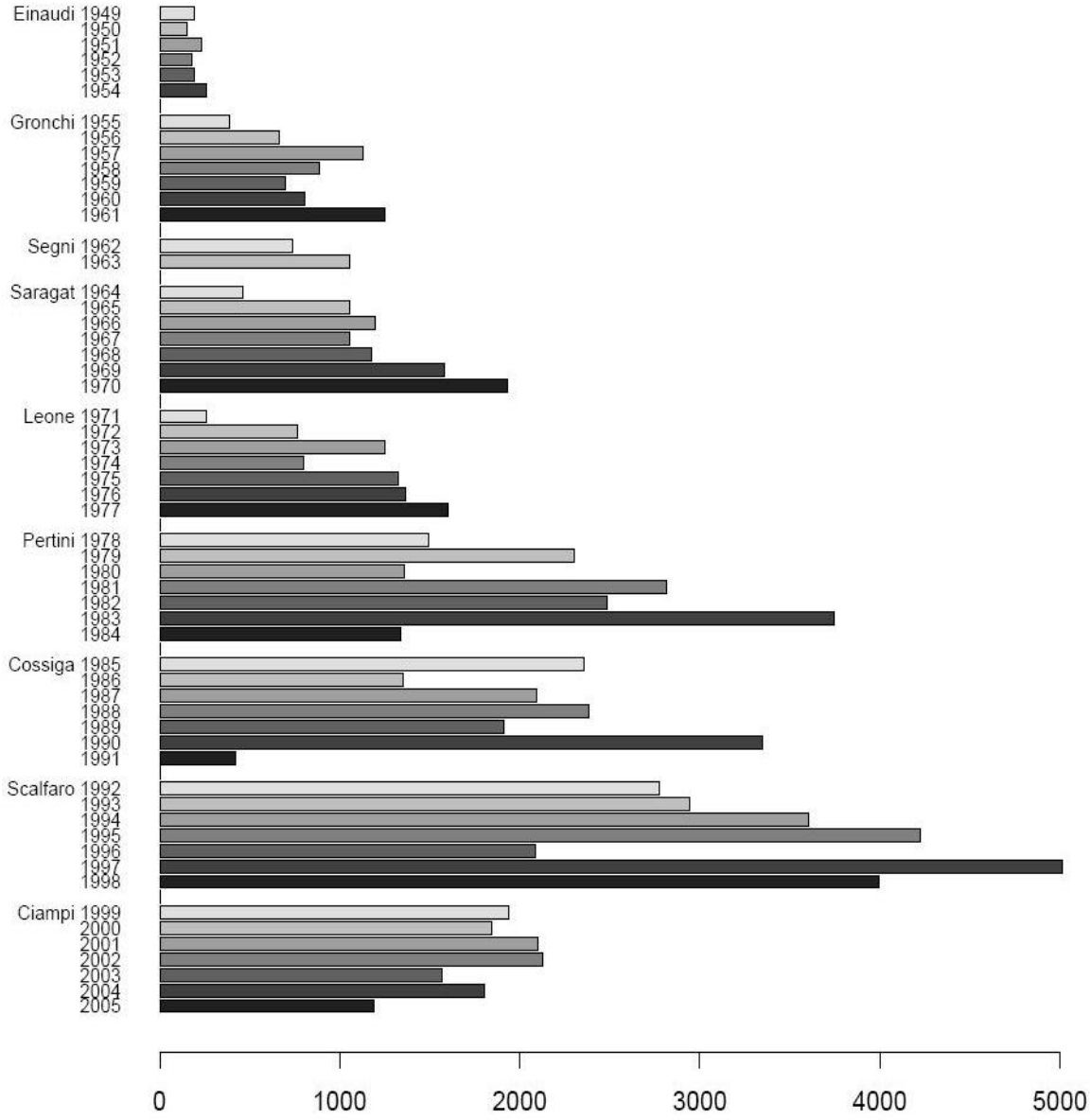


Figure 1: Lengths of presidential addresses

In order to increase the amount of information available, word forms of the corpus were turned into lemmas. Lemmatisation was conducted through a partly manual and partly automatic process associating each word-form to its corresponding lemma that has a part of speech tag. Manual disambiguation of all lemmas with context check of each occurrence was performed. Lemmatisation is more important in Italian than in other languages, reducing 'noise' (the variation in forms) and increasing the amount of information conveyed by each lemma-type. In addition, in Italian, owing to the wide range of contingent variations (masculine, feminine and plural forms, verb conjugations, clitic pronouns, etc.), lemmatization reduces the number of different types. The corpus is composed of $N = 91436$ word-tokens and 9786 word types. After lemmatization there are $V = 6003$ lemma-types. From a statistical point of view, the tokens represent the statistical units, and the lemma-types the items of the lexicon variable. The addresses and presidents are additional classification variables for the statistical units.

It is worth looking at the lemma-vocabulary ordered by decreasing frequency (Table 2), to comprehend some recurrent topics. The frequency of a word is, in fact, a rough but effective indicator of the importance of a topic in the corpus. Taking into account content lemmas only,

the common topics of the addresses correspond to the most frequent lemmas, the nouns: *anno/year* (450), *popolo/people* (328), *Italia/Italy* (316), *paese/country* (245), *pace /peace* (239), *italiano/Italian* (219), *mondo/world* (206), *cittadino/citizen* (189), *augurio/wish* (188), *libertà/freedom* (179). These words are all related to the particular circumstances of the addresses. It is also worth mentioning that the main part of the vocabulary is composed of low frequency types, 2333 lemma-types are hapax legomena (frequency equal to one), 967 are dis legomena (frequency equal to two), and only 1060 have frequency greater than nine.

Table 2
Excerpt of the lexical contingency table

Lemma	GRAM	Einaudi	Gronchi	Segni	Saragat	Leone	Pertini	Cossiga	Scalfaro	Ciampi	*Corpus*
di	PREP	92	531	164	814	620	802	1253	1689	996	*6961*
il	DET	51	211	71	335	223	564	449	903	521	*3328*
essere	V	31	116	17	249	243	692	337	1030	333	*3048*
e	CONJ	36	240	74	283	243	426	625	759	329	*3015*
a	PREP	47	185	67	222	260	450	406	751	397	*2785*
in	PREP	49	154	53	254	173	348	425	469	382	*2307*
la	DET	10	113	35	183	122	343	264	512	259	*1841*
uno	DET	12	120	29	145	167	300	304	470	250	*1797*
che	PRON	29	97	28	154	159	377	247	440	230	*1761*
avere	V	8	46	33	94	65	289	146	479	202	*1362*
per	PREP	10	96	23	97	79	199	181	323	177	*1185*
non	ADV	6	43	14	60	76	164	138	361	88	*950*
:	:	:	:	:	:	:	:	:	:	:	:
anno	N	18	31	7	43	36	66	62	132	55	*450*
:	:	:	:	:	:	:	:	:	:	:	:
popolo	N	8	26	14	20	13	120	42	54	31	*328*
io	PRON	1	14	1	10	6	189	13	79	4	*317*
Italia	N	3	14	6	35	12	37	26	93	90	*316*
mi	PRON	2	5	6	20	14	111	25	83	31	*297*
:	:	:	:	:	:	:	:	:	:	:	:
paese	N	5	14	6	44	35	37	62	22	20	*245*
pace	N	2	15	8	25	7	40	37	69	36	*239*
:	:	:	:	:	:	:	:	:	:	:	:
italiano	N	8	15	6	16	15	72	12	27	48	*219*
mondo	N	1	8	4	28	9	30	29	55	42	*206*
:	:	:	:	:	:	:	:	:	:	:	:
italiano	A	3	9	5	15	7	72	11	41	26	*189*
cittadino	N	3	11	3	20	15	7	59	47	24	*189*
augurio	N	4	2	0	22	15	15	17	93	20	*188*
libertà	N	5	16	6	23	8	20	57	23	21	*179*
giovane	N	0	0	0	7	9	95	5	31	29	*176*
problema	N	2	26	2	27	26	15	34	28	12	*172*
vita	N	3	14	3	16	9	33	22	49	23	*172*
:	:	:	:	:	:	:	:	:	:	:	:
Stato	N	0	14	1	8	10	5	32	74	13	*157*
:	:	:	:	:	:	:	:	:	:	:	:
forza	N	0	6	1	11	19	20	25	36	29	*147*
Europa	N	0	2	0	8	3	15	31	44	43	*146*
:	:	:	:	:	:	:	:	:	:	:	:
operaio	A	0	0	0	1	0	6	0	0	0	*7*
contadino	N	0	0	0	5	0	2	0	0	0	*7*
moglie	N	0	0	0	0	0	2	0	0	5	*7*
ringraziare	V	0	0	0	0	0	0	0	6	1	*7*
Risorgimento	N	0	0	0	1	0	0	0	0	6	*7*
memoria	N	1	0	1	0	0	0	0	2	3	*7*
:	:	:	:	:	:	:	:	:	:	:	:
lager	N	0	0	0	0	0	1	0	0	0	*1*
fascista	N	0	0	0	0	0	1	0	0	0	*1*
nazifascista	A	0	0	0	0	0	1	0	0	0	*1*

nazista	N	0	0	0	0	0	1	0	0	0	*1*
processuale	A	0	0	0	0	0	0	0	1	0	*1*
condivisione	N	0	0	0	0	0	0	0	0	1	*1*
identificazione	N	0	0	0	0	0	0	0	0	1	*1*

3. Distinctive Words of the Nine Presidents

In order to assess the characteristics of one president's addresses with respect to those by others, a selection of words that are used in an exclusive manner (*i.e.* they occur only in the addresses of a president and never in the others') could be a first step to understand some characteristic features. But from a non-deterministic point of view, words occurring noticeably more (or less) often in a president's address(es) than in the others' and in the corpus as a whole are useful. The traditional 'characteristic textual units' method (Lafon 1980, Lebart *et al.* 1998, Bolasco 1999) is a simple tool based on the hypergeometric model. All words which show a high probability of over-usage for an author can be considered distinctive to that author with reference to the others.

The issue of deciding whether a word appears homogeneously across the nine presidents, or if it appears mostly in one or a few authors is traditionally tackled using Fisher (χ^2) test (Casella and Berger 2002). A word which is homogeneously distributed across authors should appear in each author's subcorpus with a frequency roughly proportional to the size of the author's subcorpus. In other words, we must deal with the issue of testing the difference between observed and expected frequency distributions of a certain word across subcorpora where the expected distribution implies frequencies (probabilities) proportional to the size of the subcorpora. In Pauli and Tuzzi (2006) we discussed reasons to prefer bootstrap over both χ^2 and hypergeometric-based tests because the results of the χ^2 test are seriously affected by the size of subcorpora, and hypergeometric-based tests are time consuming. Moreover, the bootstrap approach seems not to suffer lack of robustness for low frequency words.

In the bootstrap alternative, the words in the whole corpus are resampled with replacement according to the following rule: if *x* is the sample of words (that is, *x* is the vector of 91259 words in the corpus) and $n_{\cdot 1},...,n_{\cdot 9}$ are respectively the sizes of the nine subcorpora, we resample with replacement from *x* to form a vector x^* of the same length: the first $n_{\cdot 1}$ elements of *x* are then the bootstrap resample for the first president, the second $n_{\cdot 2}$ the resample for the second president and so on. We have then a bootstrap sample of word frequencies, and so, for each iteration, we build a (lexical) contingency table (words × presidents). As a measure of the homogeneity of a row, we compute the maximum of the absolute differences between the observed (absolute) frequencies and the expected (absolute) frequencies calculated under the assumption of even distribution of words among subcorpora. In other terms, if n_i represents the frequency of the word in the whole corpus and $n_{\cdot j}$ represents the size of subcorpus *j*, and $N(=n_{\cdot\cdot})$ the size of the corpus, expected frequencies $\hat{n}_{ij}$ are given by $n_i n_{\cdot j} / N$. We compute the distance between each bootstrap distribution F^* and the expected distribution *F* by

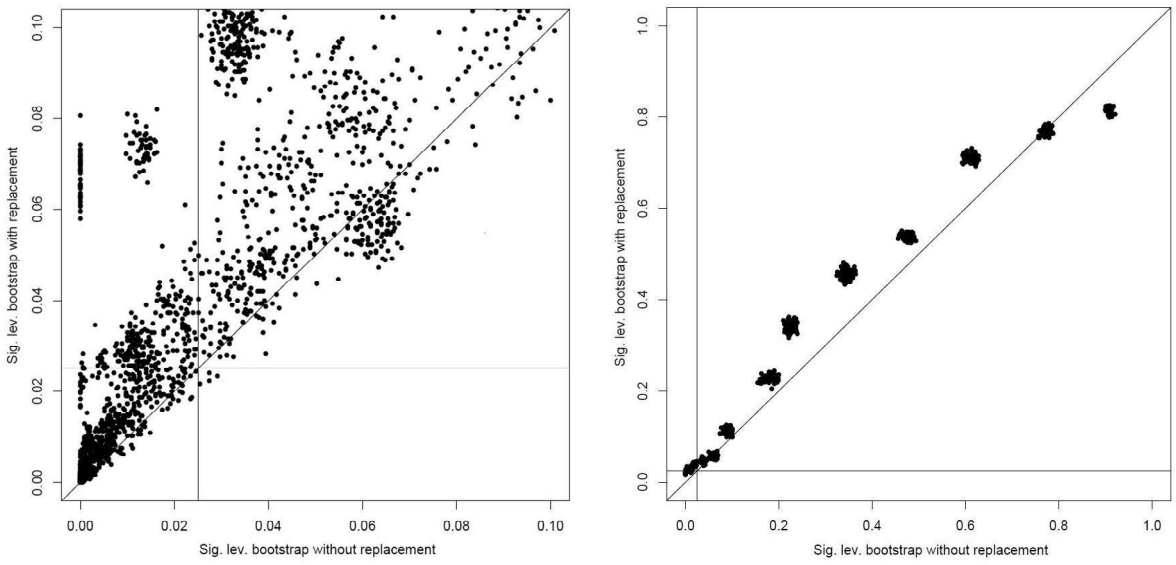
$$d(F^*, \hat{F}) = \max_i \{| n_i^* - \hat{n}_i |\}$$

and compare these with the corresponding distance between the observed distribution $\tilde{F}_{obs}$ and *F*: $d(\tilde{F}_{obs}, \hat{F})$. The bootstrap *p*-value is then given by

$$\frac{1}{B}\sum_{b=1}^{B} | \, d(F^{*b}, \hat{F}) > d(\tilde{F}_{obs}, \hat{F}) \, |$$

where B is the number of bootstrap replications. We choose resampling with replacement in order to avoid low frequency words to appear in all bootstrap samples. The resampling scheme is such that the frequency of the words in the bootstrap samples changes, that is, the row total in the (lexical) contingency table are not held fixed. This choice is, we believe, particularly appropriate to deal with low frequency words. If we resample without replacement, a word which appears once in a corpus of length 10000 would be bound to appear in each bootstrap sample and would appear in subcorpus j with a probability proportional to its size. Moreover, by resampling with replacement we limit the amount of distinctive words because the p-value is less than the one obtained by resampling without replacement (cf. Figure 2).

The occurrence of a word in an author is not a simple attribute because a subcorpus is composed of different addresses (in most cases seven), and the occurrence of a word in the subcorpus is the sum of the occurrences for each address delivered by the same president. If a word occurs a great deal more in one or a few addresses, it is erroneously considered distinctive for the president and not only for that address. Measures of the dispersion (Baayen 2001), Tuzzi and Tweedie 2000, Pauli and Tuzzi 2006) of the words within addresses are important to test if a word is well spread out over all the addresses of the same president. Only in the second case can a word detected as distinctive be actually considered a characteristic feature of the End of Year Addresses of that president. From a technical point of view, this issue is not different than that we dealt with in section 3 where the president we are investigating in depth plays the role of the corpus and the addresses play the role of the presidents. In this case we consider only bootstrap with replacement.



a) b)

Figure 2. Comparison between bootstrap p values obtained by resampling with and without replacement for words of any frequency (a), of frequency two (b).

4. Presidents' Preferred Words and Topics

In the discussion of the results, the lemmas are considered which are distinctive of one president with respect to the other eight, but which are not distinctive in the addresses by the same president. Table 3 shows the most significant lemmas of three presidents: Luigi Einaudi, Sandro Pertini and Carlo Azeglio Ciampi. The asterisks indicate that the test based on the bootstrap method and the test based on the hypergeometric distribution are significant (columns *b* and *h* respectively), and that the lemma is distinctive of one president in the comparison with the other eight. The fourth column (*o*) shows the result, in terms of *p*-value, of the bootstrap test verifying the homogeneity internal to the president's addresses. A non-significant result shows that the word is well spread-out in the seven addresses.

Table 3
Excerpt of specific words for three presidents

Einaudi				Pertini				Ciampi			
word	*b*	*h*	*o*	*word*	*b*	*h*	*o*	*word*	*b*	*h*	*o*
nuovo_A	*	*	0.982	adesso_ADV	*	*	0.119	caro_A	*	*	0.647
foriero_A	*	*	0.715	americano_A	*	*	0.277	città_N	*	*	0.020
ora_N	*	*	1.000	amico_N	*	*	0.001	come_CONG	*	*	0.652
elevare_V	*	*	0.193	animo_N	*	*	0.021	Euro_N	*	*	0.185
patria_N	*	*	0.968	anziano_N	*	*	0.111	Europa_N	*	*	0.663
confortare_V	*	*	1.000	ascoltare_V	*	*	0.016	europeo_A	*	*	0.599
volgere_V	*	*	1.000	assassinare_V	*	*	0.074	fiducia_N	*	*	0.584
tappa_N	*	*	0.758	avere_V	*		0.000	forte_A	*	*	0.439
concorde_A	*	*	0.393	carcere_N	*	*	0.636	generazione_N	*	*	0.900
borgo_N	*	*	0.016	cercare_V	*	*	0.058	identità_N	*	*	0.421
domani_N	*	*	0.907	che_PRON	*		0.255	immagine_N	*	*	0.368
fecondo_A	*	*	0.851	combattere_V	*	*	0.003	istituzione_N	*	*	0.005
anno_N	*	*	0.971	contingente_N	*	*	0.000	Italia_N	*	*	0.843
prova_N	*	*	0.951	contro_PREP	*	*	0.023	mente_N	*	*	0.791
soddisfazione_N	*	*	0.917	dannato_A	*	*	0.596	nostro_A	*	*	0.404
casolare_N	*	*	0.016	difendere_V	*	*	0.006	patria_N	*	*	0.169
proseguire_V	*	*	0.701	discorso_N	*	*	0.836	quello_A	*	*	0.147
idealmente_ADV	*	*	0.593	disoccupazione_N	*	*	0.164	secolo_N	*	*	0.002
elevazione_N	*	*	0.536	dittatura_N	*	*	0.742	Unione_N	*	*	0.807
lieto_A	*	*	0.268	domanda_N	*	*	0.114	vi_PRON	*	*	0.033
in_PREP	*	*	0.535	domani_ADV	*	*	0.535	vivere_V	*	*	0.247
sollecito_A	*	*	0.847	ebbene_CONJ	*	*	0.241	avvertire_V	*	*	0.847
via_N	*	*	0.338	essere_V	*		0.019	dare_V	*	*	0.232
apprestare_V	*	*	0.779	fame_N	*	*	0.003	governare_V	*	*	0.581
percorso_N	*	*	0.482	fare_V	*		0.072	inno_N	*	*	0.048
ancor_ADV	*	*	0.078	fianco_N	*	*	0.252	prestigio_N	*	*	0.540
tutto_PRON	*	*	0.160	funerale_N	*	*	0.060	provincia_N	*	*	0.371
voto_N	*	*	0.548	giovane_N	*	*	0.000	sogno_N	*	*	0.223
sicché_CONJ	*	*	0.242	gioventù_N	*	*	0.051	stesso_A	*	*	0.005
lecito_A	*	*	0.715	guerra_N	*	*	0.030	su_PREP	*	*	0.559
auspici_N	*	*	0.125	invece_AVV	*	*	0.993	voi_PRON	*	*	0.280
riservare_V	*	*	0.847	io_PRON	*	*	0.729	affidare_V	*	*	0.579
affetto_N	*	*	0.917	italiano_A	*	*	0.000	comunale_A	*	*	0.581
comune_A	*	*	0.849	italiano_N	*		0.020	confronto_N	*	*	0.512
ricostruzione_N	*	*	0.242	Libano_NM	*	*	0.012	disastro_N	*	*	0.006
accogliere_V	*	*	0.779	loro_A	*	*	0.019	in_PREP	*	*	0.695
muovere_V	*	*	0.125	loro_PRON	*	*	0.018	italiano_N	*	*	0.077
opera_N	*	*	0.765	mafia_N	*	*	0.099	mio_A	*	*	0.155
ideale_A	*	*	0.653	mi_PRON	*	*	0.257	ora_AVV	*	*	0.462
italiano_N	*	*	0.953	mila_A	*	*	0.333	Risorgimento_N	*	*	0.360
perseguire_V	*	*	0.242	miliardo_N	*	*	0.457	sfida_N	*	*	0.459
ognuno_PRON	*	*	0.788	mio_A	*	*	0.034	unità_N	*	*	0.029
avvenire_N	*	*	0.661	molto_A	*	*	0.079	animare_V	*	*	0.323
cammino_N	*	*	0.758	morire_V	*	*	0.900	coesione_N	*	*	0.041
sereno_A	*	*	0.701	napoletano_A	*	*	0.050	fondamentale_A	*	*	0.045
conservazione_N	*		0.125	ordigno_N	*	*	0.980	Balcani_NM	*	*	0.724
ogni_A	*	*	0.632	perché_CONG	*		0.281	buongoverno_N	*	*	0.001

pensiero_N	*	*	0.305	pidue_N	*	*	0.000	europeo_N	*	*	0.761
via_ADV	*	*	0.125	poi_AVV	*	*	0.729	spirito_N	*	*	0.227
frapporre_V	*		0.125	popolo_N	*	*	0.001	vostro_A	*	*	0.001
:	:	:	:	:	:	:	:	:	:	:	:
palpito_N	*		0.593	umanità_N	*		0.837	moglie_N	*	*	0.862
ognor_ADV	*		0.784	Spagna_NM	*	*	0.168	sessanta_NUM		*	0.045
rigoglio_N	*		0.593	sismico_A	*	*	0.001	immigrazione_N		*	0.840
ordunque_CONJ	*		0.274	Hiroshima_NM	*	*	0.002	sindaco_N		*	0.009
Trieste_NM	*		1.000	Dozier_NM	*	*	0.045	federalismo_N		*	0.840
duro_A		*	0.593	colma_A		*	0.015	cento_NUM		*	0.021
				scandalo_N		*	0.764	millennio_N		*	0.902
				40000_NUM		*	0.007	Repubblica_N		*	0.013
				Bologna_NM		*	0.039	catastrofe_N		*	0.045
				antisismico_A		*	0.015	votare_V		*	0.009
				Sambro_NM		*	0.007	Asia_N			0.045

Einaudi appears to be the spokesperson of a country which is not yet fully developed and of an archaic lexicon which is not found in his successors (*foriero/afoot*, *borgo/village*, *casolare/country house fecondo/fecund*, *palpito/palpitation*, *ognor/now*, *ricostruzione/ reconstruction*). Pertini shows colloquial and informal features (*ebbene/yet*, *io/I*, *mi/me*, *mio/my*), but also tragic themes discussed with great emphasis (*assassinare/to murder*, *carcere/jail*, *combattere/to fight*, *difendere/to defend*, *dannato/damned*, *dittatura/dictatorship*, *fame/hunger*, *funerale/funeral*, *guerra/war*, *mafia/mafia*, *morire/to die*, *ordigno/bomb*). With Ciampi, the concept of homeland finds renewed strength (*identita/identity*, *Italia/Italy*, *patria/ homeland*, *inno/anthem*, *Risorgimento/Italian Risorgimento*, *unita/unity*, *coesione/cohesion*, *spirito/spirit*, *Repubblica/Republic*), within a perspective that promotes nationalism from local autonomies (*città/city*, *provincia/province*, *comune/municipality*, *comunale/municipal*, *sindaco/mayor*) to the entire country and as far as Europe (*Europa/Europe*, *Unione/Union*, *europeo_A/ european_ADJ*, *europeo_N/european_NOUN*).

Through the presidential addresses it is possible to individuate values and reconstruct real presidential political plans. Some figures stand out clearly, while others fade in the background. Overall, Gronchi, Segni (of whom we only have two addresses) and Leone's addresses do not show clear distinctive characters. Cossiga, from his first address, seems concerned with cutting loose from the very popular figure of his predecessor, and innovating in his choice of international and institutional themes. Saragat stands out, who first clearly opens the doors to political lexicon. But it is mainly Pertini, Scalfaro and Ciampi, who show well-defined characteristic features (Bernardi and Tuzzi 2007).

5. President Napolitano's Two Addresses

Our research on the presidents' distinctive words and preferred themes ends with Ciampi, since he is, at the moment, the last president who has completed his term. Nothing can yet be said on the style of the present President of the Republic Giorgio Napolitano (Cortelazzo and Tuzzi 2006). To be able to draw conclusions on the presidents' communication styles, it is necessary to study the whole series of addresses. In fact, almost all the presidents start off in a measured manner, and then familiarize themselves with the discourse genre and the communication medium, and then show their own style and their own theme and expression preferences. To gain a view of Napolitano's two addresses, delivered in 2006 and 2007, it is appropriate to change the perspective of analysis. To place the 59 addresses delivered by the ten presidents (including Napolitano) in terms of reciprocal proximity, we decided to use the concept of distance based on lexical connection, introduced by Brunet (1988), and recently developed by

Labbé (2007), Labbé and Labbé (2001), Merriam (2002). Unlike the cited works, in this study the calculations were made on the basis of lemmas and not forms.

Given a pair of texts *A* and *B* of length N_a and N_b with $N_a \leq N_b$, the frequency F_{ib} of each type *i* in the larger text *B* is reduced according to the size of the shorter *A* in estimating the mathematical expectancy of the frequency of the type *i* in *A*,

$$\hat{F}_{ib} = F_{ib} \frac{N_a}{N_b}$$

by means of a simple proportion, hence:

$$\hat{N}_b = N_a .$$

The distance between text *A* and text *B* is obtained by:

$$\frac{\sum_{i \in V_{A \cup B}} | F_{ia} - F_{ib} |}{2N_a}$$

where *V* is the vocabulary size of text *A* and text *B*. With reference to other studies, we did not limit calculations to the types of *B* whose frequency is high enough to expect almost one in *A*. ($\hat{F}_{ib} \geq 1$), because the simple version of the procedure leads to the same results. Moreover, we did not apply the 'sliding window method' (Labbé 2007) because our addresses have lengths under 10 000 tokens and in most cases the differences in terms of sizes do not exceed a ratio of 1:8.

The two addresses of Giorgio Napolitano (Figure 3) are similar. Then, Ciampi and Cossiga's addresses are the most similar. A cluster analysis of the 59 addresses has been performed using the above distance with complete linkage (the distance between two clusters is the maximum distance between elements of each cluster) and an agglomerative hierarchical cluster algorithm. We used complete linkage since we expected to find well separated and tight (convex shaped) clusters. The dendrogram (Figure 4) shows that the individual characteristic features are more important than other factors such as time proximity, political identity and career, and professional profile. These factors play a role in determining the themes and the style of the address, but the personal trait seems to be more relevant. Einaudi's addresses form an isolated cluster relatively distant to the others and this is coherent with his aforementioned peculiarities: archaic lexicon and conciseness. Gronchi seems to inherit some elements from his predecessor in his first address (1955). The address given by Leone in 1971 and that given by Cossiga in 1991 are exceptional for their brevity (Cossiga delivered a 418-word message introductory to his resigning from office). Two clear cut seven-address clusters are formed by the addresses given by Presidents Scalfaro and Pertini. As far as President Napolitano is concerned, his two addresses share a cluster with those from Cossiga and Ciampi, which appear the nearest also according to the comparison in Fig. 3. The position in the dendrogram of the last two Saragat's addresses (1969 and 1970) points out a change with respect to the previous five. Based on a qualitative analysis, he seems to have changed his discourse practice after the 1968 Italian political events (concerning in particular the protests of students and workers against the 'establishment'). As we already said addresses by Gronchi, Segni and Leone do not show marked differences with each other.

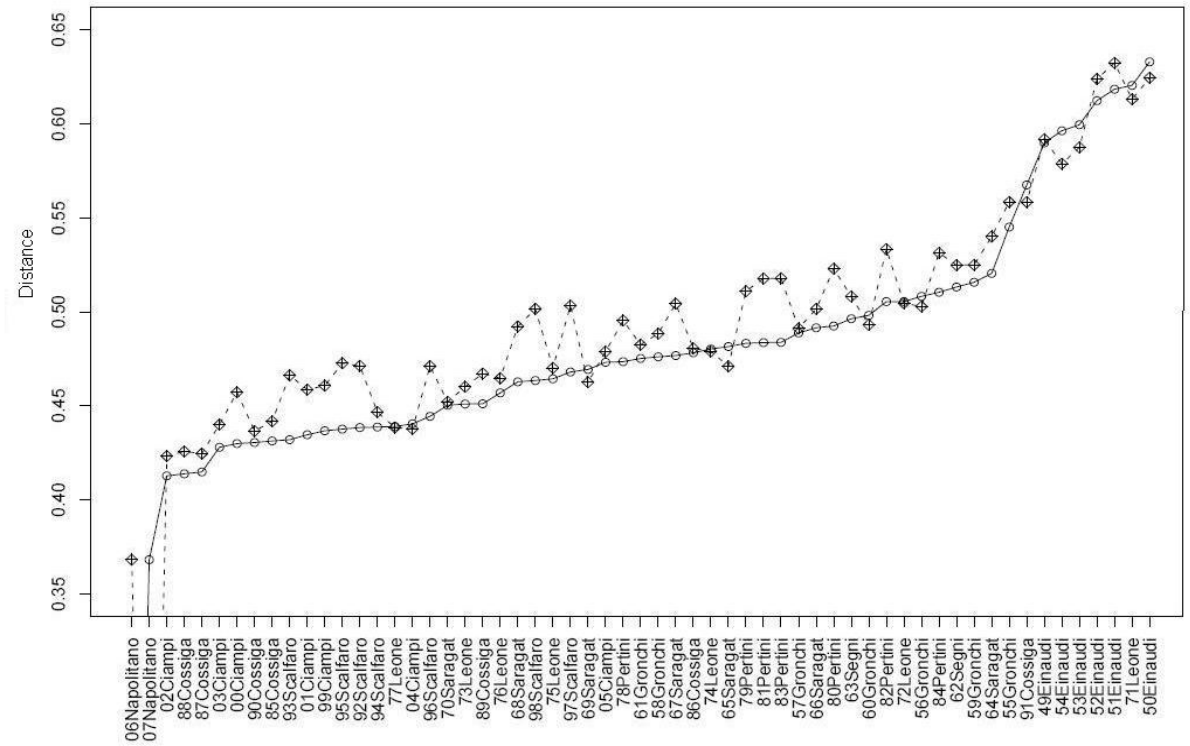


Figure 3: Intertextual distances for the two Napolitano's addresses (2006-2007)

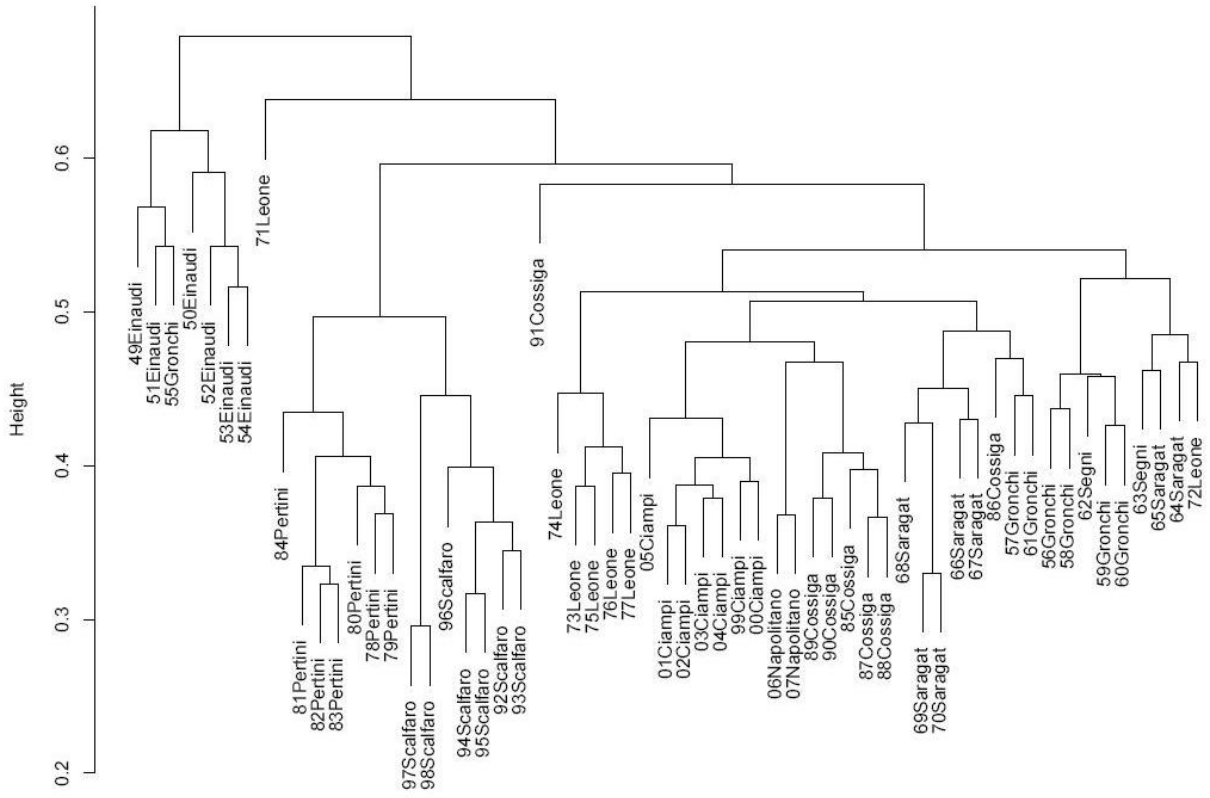


Figure 4. Dendrogram: intertextual distance: agglomerative hierarchical cluster, complete linkage

6. Conclusions

As already shown in Pauli and Tuzzi (2006) testing for homogeneity of a word among presidents is not enough to draw general conclusions about differences and distinctive features. Further analyses need to be carried out to ensure that the considered words (or lemmas) are representative of the presidents' communication styles and that peculiarities are not due to a small subset of addresses. We suggest that in order to draw sensible conclusions, these analyses should be based on a table similar to Table 3, including results of hypergeometric and bootstrap tests among the presidents ('between' perspective) and among addresses delivered by the same president ('within' perspective). Moreover, since the analysis was conducted on addresses delivered in Italian, a token-form analysis is severely limited owing to the contingent nature of some lexical choices, which do not depend on the individual's style and lexical richness. In our view a transition from the form analysis to the lemma analysis leads to more reliable results. The intertextual distance considered in section 5 proved useful to compare and contrast institutional (and political) discourses. In our case it performed well and lead to very similar results both using all words or only non-hapax words.

Contrary to what (traditionally) is expected the addresses go beyond the obvious contents of good wishes. The results show that the individual characteristic features are more important than other factors and the presidents' personal traits emerge both in the selection of the topics and in their communication choices. The End of Year Address topics and what a president decides to say (or not to say) remain unpredictable. Their position, popularity and authoritativeness allow the presidents to draw up their addresses regardless of the respect of the genre and tradition.

References

Baayen H.R. (2001). *Word Frequency Distributions. Exploring Quantitative Aspects of Lexical Structure.* Dordrecht: Kluwer Academic Pub.

Bernardi L., Tuzzi A. (2007). Parole lette con misura (statistica). In Cortelazzo M.A. and Tuzzi A. (eds), *Messaggi dal Colle: 109-134.* Venezia: Marsilio.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Roma: Carocci.

Brunet E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. Le nombre et le texte. *Revue informatique et statistique dans les sciences humaines*, Université de Liége.

Casella G., Berger R.L. (2002). *Statistical inference*. Pacific Grove: Duxbury.

Cortelazzo M.A., Tuzzi A. (2006). Il discorso di insediamento del Presidente della Repubblica Giorgio Napolitano. Lessico e retorica. In: *LId'O Lingua italiana d'oggi*: *III, 125-38* Roma: Bulzoni.

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

Cortelazzo, M.A. (1985). Dal parlato al (tra)scritto: i resoconti stenografici dei discorsi parlamentari. In Holtus G. and Radtke E. (eds), *Gesprochenes Italienisch in Geschichte und Gegenwart:. 86-118*. Tübingen: Narr.

Labbé C., Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Moliére. *Journal of Quantitative Linguistics 8, 213-231*.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics 14, 33-80*.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots,* 1: 127-65.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. Dordrecht: Kluwer-Academic Pub.

Merriam T. (2002). Intertextual Distances Between Shakespeare Plays, With Special Reference to Henry V (Verse). *Journal of Quantitative Linguistics*, 9: 261-273.

Pauli F., Tuzzi A. (2006). Identifying specific textual units of documents taken from large corpora. Comparing methods. In Viprey J.M. (ed), *JADT 2006 8es Journées internationales d'Analyse statistique des Données Textuelles*. Besançon: Presses universitaires de France-Comté, 717-728.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Tuzzi A., Tweedie F.J. (2000). The Best of Both Worlds: Combining MOCAR and MCDISP. In Rajman M. and Chappelier J.C. (eds), *JADT 2000 5es Journées internationales d'Analyse statistique des Données Textuelles*. Lausanne: EPFL ed., Vol. 1, 271-76.

Glottometrics 18, 2009, 52-68

Graphemhäufigkeiten in slawischen Sprachen:
Stetige Modelle

Emmerich Kelih, Graz

Abstract. In the article, several continuous functions will be fitted to grapheme frequencies in order to test and find the most adequate ones. They can easily be redefined in discrete forms or normalized if necessary.

Keywords: Graphem, rank-frequency, Slavic languages

0. Einleitung

Seit den Arbeiten von G. K. Zipf (1935, 1949) geht man davon aus, dass sprachliche Phänomene selbstregulierenden und selbstorganisierten Prozessen unterliegen. Diese Organisation resultiert aus einem Wechselverhältnis von Hörer- und Sprecherbedürfnissen. Ein Resultat eines versuchten Ausgleiches von Hörer- und Sprecherbedürfnissen ist die spezifische Ausprägung der Häufigkeits- und Längenstruktur von sprachlichen Entitäten.

Im vorliegenden Fall geht es um die Frage, inwiefern Graphemhäufigkeiten systematisch organisiert sind und diese durch eine stetige Funktion modelliert werden können. In letzter Zeit sind slawische Graphemhäufigkeiten (Slowenisch, Slowakisch, Russisch, Ukrainisch) durch diskrete Verteilungsmodelle (vgl. Grzybek/Kelih 2005a, 2005b; Grzybek/Kelih/Altmann 2005 u.v.m.). erfasst worden. Über die erfolgreiche Modellierung hinaus sind auch nennenswerte Zusammenhänge zwischen den iterativ bestimmten Parametern der theoretischen Verteilungen und dem zugrunde gelegten Inventarumfang aufgedeckt worden.

Diskrete und stetige Modelle sind aus mathematischer Sicht ineinander überführbar und stellen im Grunde „nur“ eine unterschiedliche Betrachtungsweise auf das gleiche Objekt, eine unterschiedliche Approximation der Wahrheit dar (vgl. Mačutek, Altmann 2007). Ebenso fakultativ ist die Alternative zwischen einer Wahrscheinlichkeitsfunktion und einfacher (nichtnormierter) Funktion. Im Folgenden wird auf der Basis von zwölf slawischen Standardsprachen eine Reihe von ausgewählten stetigen theoretischen Modellen – die bislang explizit für Graphem- bzw. Phonemhäufigkeiten diskutiert worden sind – empirisch überprüft. Von besonderer Relevanz ist dabei die Tatsache, dass für die 12 untersuchten Sprachen der jeweils gleiche Text in der entsprechenden Übersetzung (Paralleltext-Korpus) herangezogen wird, womit a priori jegliche „innere“ Datenheterogenität minimiert wird.
Folgende Probleme werden diskutiert:

1. unterschiedliche stetige Funktionen und
2. deren empirische Überprüfung an slawischen Sprachen.

1. Stetige Modelle für Graphem- und Phonemhäufigkeiten

Wie bei der Modellierung von Graphem- und Phonemhäufigkeiten üblich, werden die erhaltenen Rohdaten in eine Ranghäufigkeit transformiert. Dies ist die einfachste Art, wie man eine nominale Klassifikation in eine ordinale Klassifikation überführt. Demnach ergibt sich eine monoton abfallende Folge, die potentiell durch die in 1.1. diskutierten stetigen Funktio-

nen empirisch erfasst werden kann. Die Modellierung ist kein Selbstzweck, sondern ein Zwischenschritt, der letztendlich in eine linguistische Interpretation der erhaltenen Parameter mündet. Im idealen Fall sollte die Funktion selbst theoretisch abgeleitet werden.

Es folgt eine kurze Auswahl der wichtigsten stetigen Funktionen, die für Graphem- bzw. Phonemhäufigkeiten diskutiert worden sind.[1]

1.1. Zipfsches Gesetz

Das bekannteste und in der Linguistik sehr häufig diskutierte Modell ist das sogenannte Zipfsche Gesetz. Es hat folgende mathematische Form:

(1) $$y = ax^{-b}.$$

Diese entspricht einer Potenzkurve und besagt, dass – nunmehr konkret bezogen auf Graphemhäufigkeiten – sich die relative Veränderung zwischen den einzelnen Häufigkeiten proportional zur relativen Änderung des Ranges ändert, wie an der entsprechenden Differentialgleichung $dy/y = -(b/x)dx$ sichtbar ist. Ein derartiges „Powermodell" muss als ein grundlegendes linguistisches Modell angesehen werden. Für einen historischen Überblick über unterschiedliche Modifikationen dieses Gesetzes vgl. u.a. Rapoport (1982). Es ist anzumerken, dass es bei Zipf und seinen Nachfolgern als diskrete Folge benutzt worden ist.

1.2. Tuldava (1988)

Tuldava (1988 bzw. 1995) entwickelt im Rahmen einer Studie zu Phonemhäufigkeiten im Estnischen bzw. zu linguistischen Rangfolgen im Allgemeinen folgendes Modell:

(2) $$y = a + b\ln x$$

Es handelt sich um eine logarithmische Funktion, in der *b* einen negativen Wert einnimmt und als Korrekturfaktor für eine simple logarithmische Funktion eingeführt wird.

1.3. Unified Theory: Wimmer/Altmann (2005)

Aus dem Ansatz von Wimmer/Altmann (2005) – der Generierung von unterschiedlichen stetigen Funktionen oder Verteilungen aus einem gemeinsamen Ansatz der gesetzmäßigen Beziehung zwischen den einzelnen Häufigkeitsklassen – soll Formel 2 Wimmer/ Altmann (2005: 793) mit a0 = c, a1 = b, a2 = a3 = … = 0 herangezogen werden. So ergibt sich:

(3) $$y = ae^{cx}x^{-b} = ad^{x}x^{-b}.$$

Nach Strauss/Altmann/Best (2008) lässt sich diese Formel wiederum auf Yules species/genera-Funktion (vgl. Yule 1924) zurückführen. In jedem Fall ist man hier mit einem 3-para-

[1] Es ist an dieser Stelle kein historischer Überblick über alle jemals diskutierten stetigen Funktionen für Graphemhäufigkeiten möglich. Es sollen die nach unserer Ansicht wichtigsten und neuesten Modelle analysiert werden.

metrigen Modell konfrontiert[2].

1.4. Naranan/Balasubrahmanyans (1992a,b, 2000) Funktion

In der Diskussion von für linguistische Zwecke geeigneten Funktionen präsentieren Naranan/ Balasubrahmanyan (1992a, 1992b, 1998, 2000) im Rahmen einer Modifizierung des Zipfschen Ansatzes (unter Berücksichtigung der Modifikation von B. Mandelbrot) ein dreiparametriges stetiges Modell in der Form von

$$y = Ce^{a/x}x^{-b} \tag{4}$$

mit *C* als Normierungskonstante. Ohne eine Normierung ist *C* ein empirisch zu bestimmender Parameter. Dieses Modell lässt sich direkt aus der „Unified Theory" von Wimmer/Altmann (2005) ableiten, was aber an dieser Stelle nicht weiter diskutiert werden soll. Formel (4) ist nach unserem Wissen bislang noch nicht für Graphemhäufigkeiten getestet worden.

1.5. Altmanns Rangfunktion (1993)

In kritischer Auseinandersetzung mit einem Beitrag des finnischen Linguisten Pääkkönen, der finnische Graphemhäufigkeiten (vgl. Pääkönen 1993) publizierte, entwickelt Altmann (1993) eine Rangfolge, die im besonderen für Phonemhäufigkeiten geeignet sein soll:

$$y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} y_1 \tag{5}$$

Als stetige Funktion lässt sich das Modell in der Form

$$y = \frac{C\Gamma(b+x+1)\Gamma(a)}{\Gamma(a+x+1)\Gamma(b)} \tag{6}$$

darstellen. Dieses Modell wurde bislang für Graphem- und Phonemhäufigkeiten des Finnischen (vgl. Altmann 1993), des Deutschen, des Englischen (vgl. Best 2005) und des Marathi (vgl. Rajyashree 2008: 515) diskutiert. Die Adäquatheit soll in 2.2 für slawische Graphemhäufigkeiten überprüft werden.

1.6. Modifikation des Zipfschen Gesetzes: Popescu/Altmann/Köhler (2009)

Eine der aktuellsten mathematischen Umformulierungen des Zipfschen Ansatzes findet sich in Popescu/Altmann/Köhler (2009). Im Gegensatz zu sonstigen mathematischen Modifikationen, die in der Regel auf der Hinzufügung von Parametern beruhen, argumentieren die

[2] Hier könnte der Parameter *a* als Normierungskonstante dienen.

genannten Autoren hinsichtlich ihrer Modifizierung vor allem systemlinguistisch: Jede Rangverteilung in der Linguistik entstammt aus einer heterogenen Datengruppe und ist – so die Autoren – das Resultat des Wirkens von unterschiedlichen linguistischen Subsystemen. So wird beispielsweise bezogen auf Worthäufigkeiten argumentiert, dass der Anteil von Auto- und Synsemantika, oder der Anteil von Wortarten je nach untersuchtem Text bzw. Korpus schwanken muss. Daher hat man es mit „Mischungen" von Einheiten und Subeinheiten von zu modellierenden Teilsystemen zu tun. Diese hatte bereits Altmann (1992) postuliert. Die mathematische Konsequenz aus diesen Überlegungen ist, dass hinsichtlich der passenden Funktionen ebenfalls „Mischungen" gemacht werden müssen. Insofern wird – die genaue Ableitung findet sich in Popescu/Altmann/Köhler (2009) – folgende Kombination stetiger Funktionen als adäquat für linguistische Zwecke vorgestellt:

(7) $$y = 1 + ae^{-bx} + ce^{-dx} + \ldots .$$

In Popescu/Altmann/Köhler (2009) erweist sich das Modell als zufriedenstellend für Worthäufigkeiten in 20 unterschiedlichen Sprachen. Bereits erfolgt ist eine erste Übertragung auf Phonemhäufigkeiten (vgl. Rajyashree 2009: 516). Bei dieser Anwendung wird davon ausgegangen, dass die zugrunde gelegten Subsysteme, d.h. konkret die Konsonanten- und Vokalhäufigkeiten, eine Mischung darstellen. Eine Anpassung dieses Modells an die Phonemhäufigkeiten des Marathi zeigt zufriedenstellende Ergebnisse.

Dennoch soll in Anlehnung an (7) angenommen werden, dass für Graphem- bzw. Phonemhäufigkeiten das Konzept einer Mischung bzw. von Strata möglicherweise keine Rolle spielt. Üblicherweise ist das dann der Fall, wenn Strata „kooperativ" wirken, d.h. ein Stratum wird von dem anderen „unterstützend" begleitet. Es soll daher nur eine einzige Komponente des obigen Modells in Betracht gezogen werden:

(7a) $$y = 1 + ae^{-bx}$$

Erst eine empirische Validierung kann weitere Einsichten in diese Problematik liefern.

2. Anpassungsgüte und Anzahl von Parametern

Die Übersicht in 1.1.ff. beinhaltet sieben ausgewählte stetige Modelle. Bevor nun die empirische Datenbasis näher vorgestellt wird, sei in aller Kürze das Problem der Prüfung der Anpassungsgüte (d.h. also das der Übereinstimmung von empirischen und theoretischen Häufigkeiten) angesprochen. Als erstes Maß wird der übliche Determinationskoeffizient R^2 herangezogen. Als eine gute Übereinstimmung zwischen empirischen und theoretischen Werten wird dabei ein $R^2 > 0.85$ festgesetzt, wie dies auch in anderen empirischen Studien der quantitativen Linguistik üblich ist.

Allerdings sollte die Adäquatheit eines Modells nicht allein durch die Höhe des R^2 bewertet werden. Aus linguistischer Sicht ist es von vorrangigem Interesse, die Anzahl von Parametern bzw. deren mögliche Interpretation im Auge zu behalten. Im gegebenen Fall ist es die Anzahl von Parametern (vgl. Tab. 1). In den von uns diskutierten Modellen variiert die Anzahl zwischen zwei und vier Parametern.

Tabelle 1
Anzahl von Parametern

Modell	Parameter	Anzahl
1	a, b;	2
2	a, b;	2
3	a, d, b;	3
4	a, b, C;	3
5,6	a, b, C;	3
7	a, b, c, d;	4
7a	a, b,	2

Aus methodologischer Sicht wird am Anfang, d.h. zum Zeitpunkt des Testens von mehreren (im Grunde genommen gleichwertigen) theoretischen Modellen empfohlen, jenes Modell zu präferieren, welches über den höchsten Determinationskoeffizienten (R^2) und gleichzeitig über die geringste Anzahl von Parametern verfügt. Das bedeutet, dass man am Anfang zur groben Vereinfachung tendiert. Mit diesem Vorgehen hat man aber keineswegs die Gewähr, dass „richtige" Modell gefunden zu haben, sondern nur eine Möglichkeit, schwer behandelbare Probleme traktabel zu machen (cf. Bunge 1963, 1967, 1983a, b). Später werden alle Modelle komplexer, nicht einfacher, wobei in einem ersten Schritte eine Übereinstimmung der theoretischen Daten mit empirischen Gegebenheiten wünschenswert ist.

3. Slawische Graphemhäufigkeiten

Die genannten sieben stetigen Modelle, die bislang auf slawische Graphemhäufigkeiten nicht angewendet worden sind, sollen nunmehr an Datenmaterial aus slawischen Sprachen getestet werden. Zuvor ist die verwendete empirische Basis näher vorzustellen.

3.1. Paralleltexte als adäquate Versuchsbasis

Ein geschichtlicher Abriss der zahlreichen Untersuchungen zur Phonem- und Graphemfrequenz slawischer Sprachen kann an dieser Stelle nicht erfolgen (vgl. dazu Grzybek 2009). Hervorgehoben werden muss aber, dass bislang keine systematische Untersuchung (Modellierung) der Graphemhäufigkeit aller slawischen Standardsprachen vorliegt.

Für die von uns intendierte empirische Prüfung von stetigen Modellen wurde ein Parallel-Textkorpus des russischen Romans „Kak zakaljalas' stal'/Wie der Stahl gehärtet wurde" (1932-1934) von N. Ostrovskij erstellt. Es besteht aus dem russischen Original und elf Übersetzungen in weitere slawische Standardsprachen. Vgl. dazu Tab. 2 mit den Daten zum Bulgarischen, Kroatischen, Mazedonischen, Obersorbischen, Polnischen, Russischen, Serbischen, Slowakischen, Slowenischen, Tschechischen, Ukrainischen und Weißrussischen. Es sind zehn Kapitel (von insgesamt 18) dieses Romans analysiert worden, wobei nicht die Teilkapitel einzeln, sondern das jeweilige Gesamtkorpus untersucht wurde.

Die Untersuchung von „ganzen", semantisch abgeschlossenen Texten (= alle Teilkapitel), lässt sich damit begründen, dass so der Versuch gelingen kann, in etwa homogene Texte zu analysieren. Damit soll das Problem einer evtl. auftretenden Datenheterogenität möglichst gering gehalten werden. Dass ein Paralleltext- Korpus, d.h. Übersetzungen von Texten, herangezogen wird, bedarf allerdings einer gesonderten Begründung. Auch wenn übersetzte

Texte als eine Art „dritter Kode“, als nicht authentische Sprache – in diesem Zusammenhang wird vom Problem der „translationese“ gesprochen (vgl. Mauranen 2002) – aufgefasst werden können, ist nach unserer Meinung die Untersuchung von Paralleltexten ein guter Ausgangspunkt für eine kontrastive quantitative Analyse ausgewählter Strukturmerkmale (Graphem-Phonemhäufigkeiten, Silbenstruktur, Wortlänge, Type-Token-Ratio, Satzlänge usw.). Man kann davon ausgehen, dass die Übersetzungen aus dem Russischen keine abweichenden, „unnatürlichen“, sondern zumindest grammatikalisch korrekte, flüssig lesbare und verständliche Texte sind. Besondere Relevanz erhält ihre quantitative Analyse auch deswegen, weil man durch die Übersetzungen einen Einblick erhält, auf welche Weise die gleiche bzw. eine ähnliche „semantische Menge“ in unterschiedlichen Sprachen ausgedrückt wird und welche Konsequenzen damit in der quantitativen Struktur einhergehen.

Um auf die Graphem-Analyse, die einen ersten Schritt in die vergleichende quantitative Analyse von Paralleltexten darstellt, zurückzukommen: Als Zähleinheit wurden die jeweiligen Grapheme aus den in Referenzwerken zu slawischen Sprachen angeführten Alphabete[3] (vgl. Rehder 2006 bzw. Comrie/Corbett 1993) herangezogen. Auf eine Inkonsequenz, die sich aus dieser Art der Feststellung des Graphem- bzw. Buchstabeninventars von Sprachen ergibt, sei hier allerdings hingewiesen: Für das Polnische werden im „kanonischen“ Alphabet die Buchstabengruppen (ch, cz, dz, dź, dż, rz, sz) nicht als eigene Einheiten angeführt (sehr wohl aber die Einzelkomponenten), während im Kroatischen, Slowakischen, Tschechischen und Obersorbischen derartige Digraphen als autonome Bestandteile des Alphabets (da sie auch die jeweils eigenen Laute repräsentieren) aufgefasst werden[4].

Im Zusammenhang mit der Bestimmung des Inventarumfangs ist des Weiteren darauf zu verweisen, dass sich der von uns verwendete (vg. dazu Tab. 2) von dem „kanonischen“ Inventarumfang der untersuchten Alphabete unterscheiden kann. Es zeigt sich in einigen untersuchten Texten (= Sprachen), dass manche vom Alphabetsystem vorgegebene bzw. „erlaubte“ Einheiten schlichtweg nicht vorkommen. Dies gilt z.B. für das <ë> im Russischen, das <x>, <y> und <w> im Obersorbischen und das Slowakische <q>. Damit verkürzt sich der Inventarumfang der Sprachen um diese Einheiten (vgl. Tab. 2 mit den Angaben zu diesem „syntagmatischen“ Inventarumfang), die auch Ausgangspunkt für unsere weiterführenden Analysen sind. Mit anderen Worten: Zum Grapheminventar gehören in dieser Untersuchung nur die Buchstaben, die im Text selbst vorkamen.

Die von uns berechneten Graphemhäufigkeiten in den slawischen Sprachen Bulgarisch, Kroatisch, Mazedonisch, Obersorbisch, Polnisch, Russisch, Serbisch, Slowakisch, Slowenisch, Tschechisch, Ukrainisch und Weißrussisch finden sich – sortiert in alphabetischer Reihefolge und mit absoluten und relativen Werten versehen – im Anhang 1, 2 und 3 dieses Beitrages.

Die Stichprobengröße *N*, das heißt die Anzahl von Graphemen, ist bei der von uns verwendeten Untersuchung von besonderem Interesse, da – vermutlich erstmals – vergleichende Angaben zur Stichprobengröße eines Textes in unterschiedlichen slawischen Sprachen vorliegen. Da editionsbedingte Unterschiede bei den verwendeten Übersetzungen (d.h. die Verwendung unterschiedlicher Quelltexte, überarbeitete und erweiterte Auflagen u.ä.) mit hoher Wahrscheinlichkeit ausgeschlossen werden können (Details dazu in Kelih 2009), zeigt

[3] Es ist hier nicht der Ort, unterschiedliche Definitionen des Graphembegriffs zu diskutieren. Unter Graphem wird in der Regel eine Einheit verstanden, die auf graphematischer/orthographischer Ebene ein Phonem wiedergeben soll. Im vorliegenden Fall wird jedoch das Graphem als eine autonome Einheit des Zeichensystems einer Sprache ohne Interpunktionszeichen verstanden.

[4] Bislang scheint eine schriftlinguistische Behandlung des Themas zu fehlen. Anzunehmen ist aber, dass eine Untersuchung slawischer Graphemhäufigkeiten mit bzw. ohne Digraphen zu einer systematischen Verschiebung ausgewählter Kenngrößen führen würde. Vgl. dazu die Untersuchungen von Grzybek/Kelih/Altmann (2005) zur Entropie und Wiederholungs-Rate im Russischen mit und ohne den Buchstaben <ë>.

sich, dass eine in etwa gleiche „semantische Menge“ (d.h. der russische Originaltext) in anderen slawischen Sprachen durch eine durchaus recht unterschiedliche Anzahl von Graphemen ausgedrückt wird (vgl. die Daten zur Stichprobengröße in Tab. 2 mit alphabetischer Reihenfolge der untersuchten Sprachen).

Während beispielsweise der tschechische Text über 255880 Grapheme verfügt (das ist die geringste Anzahl innerhalb aller Sprachen), hat der obersorbische Text 297996 Grapheme, d.h. die größte Anzahl. Der russische Originaltext liegt mit 266055 Graphemen im Mittelfeld.

Tabelle 2
Verwendete Sprachen: Stichproben- (Grapheme) und Inventarumfang

Nr.	**Sprache**	**Inventarumfang k**	**Stichprobengröße N**
1	**Bulgarisch**	30	276131
2	**Kroatisch**	30	269384
3	**Mazedonisch**	31	283510
4	**Obersorbisch**	34	297996
5	**Polnisch**	32	291979
6	**Russisch**	32	266055
7	**Serbisch**	30	265344
8	**Slowenisch**	25	288871
9	**Slowakisch**	45	257795
10	**Tschechisch**	41	255880
11	**Ukrainisch**	34	264283
12	**Weißrussisch**	33	266237

Zu sehen ist, dass ein und derselbe Text – selbst bei nah verwandten Sprachen – durch eine recht unterschiedlich hohe Anzahl von Graphemen ausgedrückt wird. Auch wenn bislang vergleichende Untersuchungen zu diesem Fragekomplex (Textlänge übersetzter Texte) fehlen und in einem nächsten Schritt eine Messung der Stichproben in der Anzahl von Wörtern bzw. Lexemen notwendig erscheint, stellen diese Daten bereits einen durchaus interessanten und untersuchungswürdigen Befund dar.

Ob für die doch relativ großen Unterschiede in der Anzahl von Graphemen für den Ausdruck des gleichen Textes morphologische bzw. grammatikalische Faktoren, die in einem unmittelbaren Zusammenhang zur unterschiedlichen Ausprägung des Analytismus/Synthetismus stehen (morphologische Ausdruck von Tempus, Genus usw.), verantwortlich gemacht werden können, oder ob sie auf übersetzungsbedingte Faktoren zurückzuführen sind, werden erst weiterführende Analysen (Textlänge in Types/Tokens, Wort- und Satzlänge) zeigen können.

3.2. Anpassungsergebnisse

In einem ersten Schritt werden für jede der in Tab. 2 enthaltenen Sprachen für sieben stetige Funktionen die Determinationskoeffizienten (R^2) berechnet. Um sich jedoch nicht ausschließlich auf die Einzelergebnisse zu konzentrieren, wird für jedes Modell ein $\overline{R}^2$ berechnet. Das ist nichts anderes als der Mittelwert aller 12 berechneten R^2. Dieses Maß fungiert hier ausschließlich als Hilfsgröße, um einen Einblick in die allgemeine Validität und in die globale Adäquatheit eines Modells für alle slawischen Sprachen zu bekommen. Im Idealfall sollte ein gemeinsames Modell für slawische Sprachen gefunden werden, denn das würde eine Interpre-

tation nachhaltig erleichtern und auch linguistischen Gegebenheiten (strukturelle Ähnlichkeit und genetische Verwandtschaft) entsprechen.

Das Ergebnis der Anpassungen ist Folgendes (Details in Tab. 3): Nach dem Zipfschen Gesetz ergibt sich für die Sprachen ein $\overline{R}^2$ von 0.85. Das ist mit Abstand das „schlechteste" Abschneiden innerhalb der getesteten Funktionen. Obwohl der erhaltene Wert zeigt, dass das Modell offensichtlich in die „richtige Richtung" geht, ist es vor dem Hintergrund der anderen, äußerst positiven Ergebnisse, in summa als nicht überzeugend zu bezeichnen.

Der Vorschlag Tuldava's (1988), stattdessen eine logarithmische Funktion mit zwei lokalen Parametern zu verwenden, liefert gegenüber dem Zipfschen Gesetz insgesamt ein gutes $\overline{R}^2$ von 0.96. Die Ergebnisse liegen für elf slawische Sprachen bei $R^2 > 0{,}96$; einzig das Weißrussische erweist sich mit einem $R^2 = 0.77$ als Ausreißer. Das dritte getestete Modell aus dem Ansatz von Wimmer/Altmann (2005) zeigt – wie aufgrund der hohen Anzahl von Parametern (3) zu erwarten – ein durchgehend zufriedenstellendes Bild ($\overline{R}^2 = 0.98$). Es ist darüber hinaus für alle Sprachen gleichermaßen gut geeignet. Vor dem Hintergrund dieses Ergebnisses ist das Anpassungsergebnis der Funktion von Naranan/Balasubrahmanyans (1992a, b, 2000) beachtenswert, die ebenfalls drei Parameter umfasst. Interessanterweise schneidet diese Funktion aber etwas schlechter ($\overline{R}^2 = 0.94$) ab, als die zuvor diskutierte.

Die bislang für Graphem- und Phonemhäufigkeiten nur selten angewandte Altmannsche Rangfunktion (vgl. Altmann 1993) erweist sich für die getesteten slawischen Sprachen als gut passend, zumal ein $\overline{R}^2$ von 0.96 erreicht wird. Auch in diesem Fall zeigt sich, dass das Weißrussische ausbricht ($R^2 = 0.83$) und wiederum diese Sprache die insgesamt guten Gesamtergebnisse für die Altmannsche Rangfunktion ein wenig verzerrt.

Tabelle 3
Anpassungsergebnisse (R^2) für 12 slawische Sprachen[5]

Modell	**Bulg**	**Kro**	**Mz**	**O-Srb**	**Pl**	**Rus**	**Serb**	**Sk**	**Slo**	**Tsch**	**Ukr**	**Wru**	$\overline{R}^2$
Zipfsche Gesetz	0,89	0,85	0,89	0,88	0,82	0,84	0,86	0,84	0,82	0,79	0,84	0,87	0,85
Tuldava (1988)	0,99	0,97	0,99	0,98	0,97	0,99	0,97	0,98	0,96	0,98	0,98	0,77	0,96
Wimmer/Altmann (2005)	0,99	0,96	0,99	0,98	0,98	0,99	0,97	0,98	0,97	0,98	0,98	0,98	0,98
Naranan/Balasubrahmanyan (1992a,b, 2000)	0,95	0,93	0,96	0,97	0,94	0,94	0,94	0,94	0,95	0,93	0,93	0,90	0,94
Altmann (1993)	0,98	0,96	0,99	0,98	0,98	0,98	0,96	0,98	0,98	0,98	0,98	0,83	0,96
Popescu/Altmann/Köhler (2009) 4 Parameter	0,99	0,96	0,99	0,98	0,98	0,99	0,96	0,98	0,97	0,98	0,97	0,97	0,98
Popescu/Altmann/Köhler (2009) 2 Parameter	0,98	0,95	0,98	0,96	0,98	0,98	0,96	0,98	0,97	0,98	0,97	0,66	0,95

Abschließend ist auf die Funktion von Popescu/Altmann/Köhler (2009) einzugehen, die als überlagerte Funktion (mit zwei Komponenten) über vier Parameter verfügt und dementspre-

[5] Für die Sprachen werden in den Tabellen und den Abbildungen folgende Abkürzungen eingeführt: Bulg = Bulgarisch, Kro = Kroatisch, Mz = Mazedonisch, O-Srb = Obersorbisch, Pl = Polnisch, Rus = Russisch, Serb = Serbisch, Sk = Slowakisch, Slo = Slowenisch, Tsch = Tschechisch, Ukr = Ukrainisch, Wru = Weißrussisch.

chend für alle von uns getesteten slawischen Sprachen als durchgehend geeignetes Modell zu ($\overline{R}^2 = 0.98$) bezeichnen ist.

Hervorgehoben werden muss aber an dieser Stelle folgender Befund: Für die von uns untersuchten Graphemhäufigkeiten scheint keine „gemischte Funktion“ notwendig zu sein, da bereits eine einzige Komponente dieser Funktion (Modell 7a) zu guten Ergebnissen führt. Das $\overline{R}^2$ beträgt in diesem Fall 0.95. Einzig und allein das Weißrussische ($R^2 = 0.66$) bricht wiederum aus diesem Gesamtbild aus. Es ist aber zufriedenstellend durch das vierparameterige Modell von Popescu/Altmann/Köhler (2009) zu erfassen.

3.3. Weißrussisch als Sonderfall?

Der Befund, dass sich die weißrussischen Graphemhäufigkeiten vor dem Hintergrund der anderen slawischen Sprachen zufriedenstellend nur durch komplexe Modelle erfassen lassen, während alle anderen Sprachen mit „einfachen“ zweiparametrigen Modellen hinreichend beschreibbar sind, bedarf einer weiteren empirischen Überprüfung bzw. im Fall eines neuerlich sich bestätigenden Abweichens einer theoretischen Begründung.

Ein möglicher Grund für diese Abweichung könnte darin liegen, dass es sich bei dem von uns untersuchten Text um eine „schlechte“ weißrussische Übersetzung handelt, die vielleicht einen älteren, stilistisch „markierten“ Sprachzustand der 50er Jahre des 20. Jahrhunderts wiedergibt. Diese Vermutung kann allerdings empirisch mit Hilfe von Tests weiterer weißrussischer Texte widerlegt werden. Zu diesem Zweck wird ein zeitgenössischer Prosa-Text (Janka Bryl’: Galja) herangezogen (im Folgenden „Kontroll- Text“), in dem die Graphemhäufigkeiten bestimmt und in Ranghäufigkeiten transformiert werden. In Tab. 4 finden sich die Anpassungsergebnisse für alle sieben getesteten Funktionen.

Tabelle 4
Anpassungsergebnisse für Weißrussisch („Kontroll-Text)

Modell	R^2
Zipfsches Gesetz	0.82
Tuldava (1988)	0.76
Wimmer/Altmann (2005)	0.97
Naranan/Balasubrahmanyan (1992a,b, 2000)	0.89
Altmann (1993)	0.87
Popescu/Altmann/Köhler (2009) 4 Parameter	0.98
Popescu/Altmann/Köhler (2009) 2 Parameter	0.66

Deutlich wird, dass es hinsichtlich der Anpassungsgüte zwischen dem weißrussischen Text „Wie der Stahl gehärtet wurde“ und einem zeitgenössischen Prosa-Text keine großen Unterschiede gibt. Es zeigt sich das bereits in Tab. 3 erfasste Gesamtbild: Das Weißrussische weicht vom Verhalten der anderen slawischen Sprachen ab. Weder das Zipfsche Gesetz, das Modell von Tuldava (1988) noch die zweiparametrige Funktion von Popescu/Altmann/Köhler (2009) sind geeignet, da in keinem Fall ein $R^2 > 0.85$ erreicht wird. Im Grunde genommen kommen nur komplexere Modelle (mit mehren Parametern) in Frage. Hier sticht besonders

das vierparametrige Modell von Popescu/Altmann/Köhler (2009) ins Auge, welches über den besten R^2-Wert (0.98) verfügt, während das zweiparametrige Modell nur ein $R^2 = 0.66$ erreicht.

Es ergibt sich demnach folgender Befund: Erstens zeigen sowohl die weißrussischen Übersetzung von „Wie der Stahl gehärtet wurde“ als auch der zu Kontrollzwecken eingeführte zeitgenössische Text, der keine Übersetzung ist, die in etwa gleich gute Anpassungsgüte: Im Fall des „Stahl-Textes“ ein R^2 von 0.97 und für den Kontrolltext ein $R^2 = 0.99$. Man kann also mit aller gebotenen Vorsicht die „Qualität“ (Alter des Textes, Übersetzung) als Faktor für die „schlechte“ Anpassungsgüte zweiparametriger Modelle ausschließen. Darüber hinaus zeigen beide Texte ein in etwa ähnliches Modell-Verhalten.

Zweitens muss an dieser Stelle neuerlich auf die Popescu/Altmann/Köhler (2009)-Konzeption hingewiesen werden, die bewusst eine „Mischung“ von theoretischen Modellen zulässt, da linguistische Daten (und eben auch Graphemhäufigkeiten) als Gesamtmenge heterogener Subeinheiten angesehen werden. Insofern muss das Weißrussische einstweilen als Sprache mit einem komplexen und „unruhigen“ Verhalten hinsichtlich der Graphemhäufigkeiten bezeichnet werden, wobei die inhaltliche Begründung einstweilen offen gelassen werden muss.

Das in der Tat für die weißrussischen Graphemhäufigkeiten ein komplexes Modell notwendig ist, lässt sich folgendermaßen demonstrieren: Es wird zu Testzwecken eine 6-parametrige Popescu/Altmann/Köhler (2009)- Funktion konstruiert, indem angenommen wird, dass den Daten eine dreifache „Mischung“ zugrunde liegt:

(7b) $$y = 1 + ae^{-bx} + ce^{-dx} + ge^{-hx}.$$

Mit diesem empirischen „Ausschlussverfahren“ kann man tentativ erforschen, wie viele unterschiedliche „Schichten“ den weißrussischen Graphemhäufigkeiten zu Grunde liegen. Es ergibt sich ein $R^2 = 0.9757$ mit folgenden Parametern: $a = 359460$, $b = 3.62$, $c = 16626$, $d = 0.057$, $g = 659525$ und $h = 3.62$. Besondere Aufmerksamkeit gilt hier der Ausprägung der einzelnen Exponenten. Es zeigt sich, dass die Parameter b und h den gleichen Wert[6] einnehmen, woraus geschlossen werden kann, dass für das Weißrussische nicht drei Schichtungen (Mischungsebenen) anzunehmen sind, sondern den weißrussischen Graphemhäufigkeiten in der Tat zwei unterschiedliche „Komponenten“ zugrunde liegen.

Eine inhaltliche Begründung für dieses Phänomen könnte folgendermaßen aussehen: Das Weißrussische hat bezogen auf alle anderen slawischen Sprachen einen auffällig hohen Anteil von Graphemen in der ersten Ranghäufigkeitsklasse (p_1). Er beträgt 0.162 (es ist dies der Vokal *a*; siehe dazu Anhang 2), während das p_1 in dem direkt vergleichbaren und eng verwandten Russischen und Ukrainischen nicht mehr als 0.1064 ausmacht. Da keine sprachhistorischen oder phonologischen Gründe für diese großen Unterschiede ausgemacht werden können (es sind alles drei ostslawische Sprachen mit einem ähnlichem Vokalsystem und einer starken Palatalisierungskorrelation), ist das hohe p_1 im Weißrussischen vermutlich nur durch schriftlinguistische Faktoren zu erklären.

Das Weißrussische verfügt über ein äußerst phonetisches Alphabet und so wird beispielsweise das Akanne, d.h. die Reduktion von unbetonten /o/ – im Gegensatz zum Russischen – auch im Schriftbild als <a> wiedergegeben. Eine weitere „Besonderheit“ des Weißrussischen scheint auch darin zu liegen, dass der „Sprung“ zwischen der ersten und der zweiten Häufigkeitsklasse sehr hoch ist: Während der Buchstabe <a> 16,2% der Gesamtverteilung einnimmt, hat die zweite Ranghäufigkeit (der Buchstabe <n>) nicht mehr als 5,88%. Ein der-

[6] Genau das gleiche Ergebnis zeigt sich auch für den von uns nachträglich zu Kontrollzwecken eingeführten zeitgenössischen Prosa- Text. Auch in diesem Fall gilt $b = h$. ($b = 3.26$ und $h = 3.26$).

artig großer „Sprung“ zwischen der ersten und der zweiten Häufigkeitsklasse ist in den anderen slawischen Sprachen ebenfalls nicht zu beobachten. Möglicherweise liegt in diesen beiden Faktoren der Grund für die nicht überzeugende theoretische Modellierung durch zweiparametrige stetige Funktionen, weshalb auf der Basis der gegenwärtigen Befunde angenommen werden kann, dass die graphematische Ebene, sofern sie durch ähnliche Grade der Abweichung vom phonologischen System ausgezeichnet ist – das gilt für alle hier untersuchten slawischen Sprachen außer dem Weißrussischen – durch „einfache Modelle“ erfasst werden kann. Gleichzeitig bedeutet dies aber, dass das weißrussische Graphemsystem aufgrund seiner phonetischen Schreibweise dazu tendiert, einzelne Grapheme überzubelasten, denn der in den Texten weitaus häufigste Buchstabe <a> gibt sowohl ein zugrundeliegendes Phonem /o/ als auch das Phonem /a/ wieder. Man hat es also mit einer hohen orthographischen Ungewissheit in Bezug auf einzelne Grapheme zu tun. Durch diese partielle Überbelastung gestaltet sich die Häufigkeitskurve insgesamt komplexer, was sich an der Brauchbarkeit des vierparameterigen Modells von Popescu/Altmann/Köhler (2009) bestätigt hat.

Es kann angesichts dieser Befunde daher nur dafür plädiert werden, in Zukunft sowohl die Phonemhäufigkeiten dieser Sprachen zu untersuchen, als auch die Graphem-Phonem-Relation als einen Einflussfaktor in Betracht zu ziehen. Hierzu müssen allerdings erst systematische Studien zur Graphem- Phonemhäufigkeit und dem Graphem-Phonem-Verhältnis in allen slawischen Sprachen erfolgen.

3.4. Abschließende Diskussion

Nach der ausführlichen Diskussion der Problematik des Weißrussischen, welches als „Ausreißer“ in Erscheinung getreten ist, soll nun abschließend auf die eigentliche zentrale Fragestellung dieses Beitrags eingegangen werden: Welches der sieben von uns getesteten Modelle kann nun als das adäquateste bezeichnet werden? Die bislang präsentierten R^2-Werte zeigen folgende Tendenz: Es muss – vor dem Hintergrund der übrigen glänzenden Ergebnisse – eigentlich nur das Zipfsche Gesetz als unpassend ausgeschlossen werden. Alle anderen diskutierten Modelle sind im Grunde genommen gleich gut geeignet (vgl. Tab. 3), slawische Graphemhäufigkeiten zu modellieren. Lediglich für das Weißrussische müssen begründet „komplexe“ vierparametrige Modelle herangezogen werden.

Es muss jedoch nicht allein die R^2-Werte als Gradmesser für die Validität eines Modells ausschlaggebend sein, sondern auch – aufgrund der anfänglichen Simplifizierung (s.o.) – die Anzahl der Parameter. Unter Berücksichtigung dieses Kriteriums ergibt sich, dass das vierparametrige Modell von Popescu/Altmann/Köhler (2009), die Funktion von Naranan/ Balasubrahmanyan (1992a,b, 2000) und die Altmannsche Rangfunktion (1993) aufgrund der hohen Anzahl von Parametern von der weiteren Diskussion ausgeschlossen werden können. Somit sind für elf slawische Sprachen – außer dem Weißrussischen – zwei Modelle hervorzuheben, die lediglich über zwei Parameter und dabei über ähnlich gute Anpassungsergebnisse verfügen: Es ist dies das Modell von Tuldava (1988) und die zweiparametrige Funktion von Popescu/Altmann/Köhler (2009).

Damit zeigt sich, dass für die slawischen Sprachen keine gemeinsame Funktion für die Modellierung der Graphemhäufigkeiten angesetzt werden kann. Die Frage, welches der beiden Modelle insgesamt adäquater ist, muss einstweilen unbeantwortet bleiben und wird in Zukunft zu diskutieren sein (vgl. Kelih 2009b), wobei weitere Kriterien und Charakteristika eingeführt werden sollten, die Auskunft über die globale „Richtigkeit“ eines Modells geben können. Zu denken ist an die Interpretierbarkeit der Parameter bzw. an eine Diskussion der globalen statistischen Kenngrößen der empirischen Rangverteilungen (Wiederholungsrate, Entropie, Mittelwert, Ord’sche I und Ord’sche S) und vor allem an potentielle Wechselbezie-

hungen zwischen den Parametern aus den theoretischen Funktionen und empirischen Kenngrößen.

4. Zusammenfassung

Die Ergebnisse der vorliegenden Studien lassen sich folgendermaßen zusammenfassen:

1. Die Graphemhäufigkeiten in 12 slawischen Sprachen aus einem Paralleltext-Korpus können durch unterschiedliche stetige Funktionen adäquat erfasst werden. Es zeigt sich ein systematisches Häufigkeitsverhalten innerhalb der slawischen Sprachen.
2. Als Kriterium für die Selektion von – im Grunde genommenen theoretisch gleichwertigen Modellen (stetige Funktionen) – werden zwei Charakteristika vorgeschlagen: 1. die Anpassungsgüte in Form des R^2 und 2. eine möglichst geringe Anzahl von theoretischen Parametern. Die Kombination dieser zwei Charakteristika lässt eine intersubjektiv nachprüfbare Selektion von Modellen zu. Im konkreten Fall zeigt sich, dass eine von Popescu/Altmann/Köhler (2009) vorgeschlagene zweiparametrige Funktion und das Modell von Tuldava (1988) geeignet sind, um slawische Sprachen zu erfassen.
3. Das abweichende Verhalten des Weißrussischen ist für die weitere Theoriebildung von besonderem Interesse. Die Tatsache, dass die Graphemhäufigkeiten dieser Sprache im Grunde genommen nur durch komplexe Funktionen mit drei bzw. vier Parametern erfasst werden können, deutet darauf hin, dass man es hier mit einer bestimmten Art der Datenmischung bzw. mit einer Überlappung von Häufigkeitsstrukturen zu tun. Das konnte anhand des Verhaltenes der Parameter aus dem zwei-, vier- und sechsparametrigen Ansatz von Popescu/Altmann/Köhler (2009) gezeigt werden. Zu begründen ist dies möglicherweise mit schriftlinguistischen Faktoren (phonetisches Prinzip der Orthographie). Man kann jedoch auch annehmen, dass sich das Weißrussische auf dem Weg aus einem alten Gleichgewicht hin zu einem neuen Attraktor bewegt.

Literatur

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 105- 120.*

Altmann, G. (1993). Phoneme counts. Marginal remarks to Pääkkönen's article. *Glottometrika 14, 54- 68.*

Altmann, G., Lehfeldt, W. (1980). *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer. [= Quantitative Linguistics, 7]

Best, K.-H. (2005). Buchstabenhäufgkeiten im Deutschen und Englischen. *Naukovij visnik Černives'koho universitetu, vypusk (Hermans'ka filolohija) 231, 119-127.*

Bunge, M. (1963). *The myth of simplicity*. Englewood Cliffs, N.J.: Prentice-Hall.

Bunge, M. (1983a). *Exploring the world*. Dordrecht: Reidel.

Bunge, M. (1983b). *Understanding the world*. Englewood Cliffs, N.J., Prentice- Hall.

Bunge, M. (1967). *Scientific research I- II.* Berlin: Springer.

Comrie, B.,**Corbett, G.G.** (eds.) (1993): *The Slavonic languages.* London/New York: Routledge.

Grzybek, P. (2009): Phonem- und Graphemhäufigkeiten in slawischen Sprachen. In: Berger, T., Gutschmidt, K., Kempgen, S., Kosta, P. (eds.), *Handbuch Slavische Sprachen.* Berlin, New York: de Gruyter. [im Druck]

Grzybek, P., Kelih, E. (2005a). Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph. In: Altmann, G., Levickij, V., Perebyjnis, V. (eds.), *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics: 159- 179*. Černovci: Ruta.

Grzybek, P.. Kelih, E. (2005b). Towards a general model of grapheme frequencies in Slavic languages. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages: 73- 87*. Bratislava: Veda.

Grzybek, P., Kelih, E. (2005c). Häufigkeiten von Buchstaben/Graphemen/Phonemen: Konvergenzen des Rangierungsverhaltens. *Glottometrics 9, 62-73.*

Grzybek, P., Kelih, E., Altmann, G. (2004). Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie 32, 25-54.*

Grzybek, P., Kelih, E., Altmann, G. (2005). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das ‘ë’. *Anzeiger für Slavische Philologie 33, 117-140.*

Kelih, E. (2008). Modelling polysemy in different languages: A continuous approach. *Glottometrics 16, 46 -56.*

Kelih, E. (2009a): Slawische Parallel- Korpora: Projektvorstellung. [in Arbeit]

Kelih, E. (2009b): Slawische Graphemhäufigkeiten: Ein empirischer Regelkreis. [in Arbeit]

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

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

Mauranen, A. (2002): Will ’translationese’ ruin a contrastive study. *Languages in Contrast, 2(2), 161-185.*

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

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

Naranan, S., Balasubrahmanyan, V.K. (1998). Models for power law relations in linguistics and information science. J*ournal of Quantitative Linguistics 5 (1-2), 35- 61*.

Naranan, S., Balasubrahmanyan, V.K. (2000). Information theory and algorithmic complexity: Applications to linguistic discourses and DNA sequences as complex systems. *Journal of Quantitative Linguistics 7, 129- 183.*

Pääkkönen, M. (1993). Graphemes and Context: Statistical data on the graphology of standard Finnish. *Glottometrika 14, 1- 53.*

Popescu, I.- I.. Altmann, G., Köhler, R. (2009). Zipf´s law – another view. [im Druck]

Rajyashree, K.S. (2008): The Phoneme- Grapheme Correspondence in Marathi. In: Altmann, G.; Zadorozhna, I., Matskulyak, J. (2008): (eds.): *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th Anniversary of Professor V. Levicikij: 503-517.* Chernivtsi: Books – XXI.

Rapoport, A. (1982). Zipf’s law re-visited. In: Guiter, H., Arapov, M. V. (eds), *Studies on Zipf’s law: 1- 28.* Bochum: Brockmeyer [= Quantitative Linguistics, Vol. 16].

Rehder, P. (ed.) (1998). *Einführung in die slavischen Sprachen. (Mit einer Einführung in die Balkanphilogie).* Darmstadt: Wissenschaftliche Buchgesellschaft.

Strauss, U., Altmann, G., Best, K.-H. (2008). Phoneme frequency. http://lql.uni-trier.de (18.12.2008)

Tuldava, J. (1988). Opyt kvantitativnogo analiza sistemy fonem estonskogo jazyka, in: *Učenye zapiski Tartuskogo gosudarstvennogo universiteta 838, 120-133.* [engl. Über-

setzung in: Tuldava, Ju. (1995): *Methods in Quantitative Linguistics*. Trier: Wissenschaftlicher Verlag, 161- 187]

Wimmer, G., Altmann, G. (2005): Unified derivation of some linguistic laws. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *791-807.*

Yule, G.U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London Biological Sciences 213, 21- 87.*

Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology.* Boston: Hougthon Mifflin Company. [Neuauflage in Zipf, G.K. (1965), Cambridge/Massachusetts: M.I.T. Press]

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* Cambridge/Massachusetts. [Reprint in Zipf, G.K. (1972), New York: Hafner Publishing Company]

Anhang[7] 1: Rohdaten für Slowenisch, Serbisch, Kroatisch, Bulgarisch

Slo	abs.	rel.	Serb	abs.	rel.	Kro	abs.	rel.	Bulg	abs.	rel.	Mz	abs.	rel.
a	30849	0,1068	а	32507	0,1225	a	32444	0,1204	а	36841	0,1334	а	40232	0,1419
b	5336	0,0185	б	3797	0,0143	b	3808	0,0141	б	4344	0,0157	б	4360	0,0154
c	1967	0,0068	в	9949	0,0375	c	2258	0,0084	в	12224	0,0443	в	11613	0,0410
č	4429	0,0153	г	4732	0,0178	č	3075	0,0114	г	4770	0,0173	г	6127	0,0216
d	9167	0,0317	д	9929	0,0374	ć	2225	0,0083	д	9197	0,0333	д	10591	0,0374
e	29708	0,1028	ђ	649	0,0024	d	9885	0,0367	е	24724	0,0895	ѓ	303	0,0011
f	230	0,0008	е	24709	0,0931	dž	55	0,0002	ж	2197	0,0080	е	28420	0,1002
g	4755	0,0165	ж	1832	0,0069	đ	637	0,0024	з	6197	0,0224	ж	1798	0,0063
h	2923	0,0101	з	5015	0,0189	e	24320	0,0903	и	21644	0,0784	з	5191	0,0183
i	26129	0,0905	и	23473	0,0885	f	241	0,0009	й	1956	0,0071	s	66	0,0002
j	14139	0,0489	j	7794	0,0294	g	4688	0,0174	к	11329	0,0410	и	20985	0,0740
k	11569	0,0400	к	9661	0,0364	h	1665	0,0062	л	8542	0,0309	j	5219	0,0184
l	15921	0,0551	л	7958	0,0300	i	24952	0,0926	м	7339	0,0266	к	10640	0,0375
m	8753	0,0303	љ	1512	0,0057	j	10237	0,0380	н	17133	0,0620	л	7815	0,0276
n	17175	0,0595	м	9163	0,0345	k	9741	0,0362	о	23098	0,0836	љ	171	0,0006
o	25886	0,0896	н	13168	0,0496	l	7779	0,0289	п	7950	0,0288	м	7123	0,0251
p	10029	0,0347	њ	1703	0,0064	lj	1709	0,0063	р	13867	0,0502	н	17111	0,0604
r	15045	0,0521	о	25823	0,0973	m	9139	0,0339	с	13394	0,0485	њ	803	0,0028
s	14144	0,0490	п	8296	0,0313	n	13215	0,0491	т	19535	0,0707	о	30122	0,1062
š	3054	0,0106	р	13332	0,0502	nj	1769	0,0066	у	4554	0,0165	п	7753	0,0273
t	12402	0,0429	с	12728	0,0480	o	25820	0,0958	ф	362	0,0013	р	13634	0,0481
u	5515	0,0191	т	11453	0,0432	p	8384	0,0311	х	2681	0,0097	с	13152	0,0464
v	11412	0,0395	ћ	2194	0,0083	r	13457	0,0500	ц	1464	0,0053	т	20793	0,0733
z	6441	0,0223	у	12888	0,0486	s	12759	0,0474	ч	4035	0,0146	ќ	1540	0,0054
ž	1893	0,0066	ф	278	0,0010	š	3768	0,0140	ш	3220	0,0117	у	6327	0,0223
			х	1592	0,0060	t	11581	0,0430	щ	1936	0,0070	ф	563	0,0020
			ц	2239	0,0084	u	12958	0,0481	ъ	5633	0,0204	х	365	0,0013

[7] Im Gegensatz zur der Anordnung der Sprachen im Fließtext nach dem Alphabet, werden hier die Sprachen nach Gruppen nach der traditionellen Einteilung in Süd-, Ost- und Westslawisch angeführt.

			ч	3004	0,0113	v	9958	0,0370	ь	336	0,0012	ц	2015	0,0071
			џ	77	0,0003	z	5047	0,0187	ю	320	0,0012	ч	3203	0,0113
			ш	3889	0,0147	ž	1810	0,0067	я	5309	0,0192	џ	35	0,0001
												ш	5440	0,0192
ges.	288871	1	ges.	265344	1	ges.	269384	1	ges.	276131	1	ges.	283510	1

Anhang 2: Rohdaten für Russisch, Ukrainisch und Weißrussisch

Rus.	abs.	rel.	Ukr.	abs.	rel.	Wru.	abs.	rel.	Wru.[8]	abs.	rel.
а	23509	0,0884	а	22419	0,0848	а	43224	0,162	а	4159	0,1580
б	4498	0,0169	б	4618	0,0175	б	4602	0,017	б	1559	0,0592
в	12693	0,0477	в	16868	0,0638	в	8431	0,032	в	1270	0,0482
г	5026	0,0189	г	4759	0,018	г	5363	0,02	г	1172	0,0445
д	8147	0,0306	г	1	0	д	8935	0,034	д	1113	0,0423
е	21205	0,0797	д	8871	0,0336	е	9765	0,037	е	1073	0,0408
ё	0	0	е	11566	0,0438	ё	1573	0,006	ё	1058	0,0402
ж	2667	0,01	є	878	0,0033	ж	2016	0,008	ж	1057	0,0402
з	5045	0,019	ж	2101	0,0079	з	8262	0,031	з	1049	0,0398
и	17140	0,0644	з	6693	0,0253	і	12899	0,048	і	1024	0,0389
й	3288	0,0124	и	17958	0,0679	й	2551	0,01	й	989	0,0376
к	10004	0,0376	і	14123	0,0534	к	10603	0,04	к	978	0,0371
л	13265	0,0499	ї	1430	0,0054	л	10528	0,04	л	916	0,0348
м	7834	0,0294	й	3952	0,015	м	7508	0,028	м	868	0,0330
н	16143	0,0607	к	9926	0,0376	н	15656	0,059	н	814	0,0309
о	28305	0,1064	л	10339	0,0391	о	10165	0,038	о	810	0,0308
п	7733	0,0291	м	7542	0,0285	п	8384	0,032	п	729	0,0277
р	13103	0,0492	н	15985	0,0605	р	12458	0,047	р	651	0,0247
с	13980	0,0525	о	25494	0,0965	с	10776	0,041	с	644	0,0245
т	14868	0,0559	п	8327	0,0315	т	8886	0,033	т	558	0,0212
у	8396	0,0316	р	11835	0,0448	у	9457	0,036	у	555	0,0211
ф	312	0,0012	с	10521	0,0398	ў	7303	0,027	ў	441	0,0168
х	2506	0,0094	т	12146	0,046	ф	301	0,001	ф	441	0,0168
ц	1098	0,0041	у	9811	0,0371	х	3129	0,012	х	438	0,0166
ч	3679	0,0138	ф	242	0,0009	ц	6063	0,023	ц	418	0,0159
ш	2859	0,0107	х	3038	0,0115	ч	4366	0,016	ч	360	0,0137
щ	971	0,0036	ц	1937	0,0073	ш	4035	0,015	ш	313	0,0119
ъ	59	0,0002	ч	4215	0,0159	ы	11631	0,044	ы	246	0,0093
ы	5191	0,0195	ш	2963	0,0112	ь	3173	0,012	ь	237	0,0090
ь	4957	0,0186	щ	1340	0,0051	э	2128	0,008	э	197	0,0075
э	539	0,002	ю	2486	0,0094	ю	1813	0,007	ю	167	0,0063
ю	1556	0,0058	я	5640	0,0213	я	10127	0,038	я	12	0,0005
я	5479	0,0206	ь	3977	0,015	ъ	126	0,0005	’	10	0,0004
			’	282	0,0011						
ges.	266055	1	ges.	264283	1	ges.	266237	1	ges.	26326	1

[8] Dies sind die Angaben zum „Kontrolltext“ von Janka Bryl’: „Galja“. Vgl. http://www.belarus-misc.org/writer/halya1-both.htm.

Anhang 3: Rohdaten für Tschechisch, Slowakisch, Polnisch und Obersorbisch

Tsch	abs.	rel.	Sk	abs.	rel.	Pl	abs.	rel.	O-Srb	abs.	rel.
a	19595	0,0766	a	26490	0,1028	a	26718	0,0915	a	29440	0,0988
á	5229	0,0204	á	4267	0,0166	ą	3714	0,0127	b	5341	0,0179
b	4103	0,0160	ä	186	0,0007	b	4361	0,0149	c	2867	0,0096
c	3169	0,0124	b	4282	0,0166	c	12120	0,0415	č	2888	0,0097
č	2583	0,0101	c	2801	0,0109	ć	1220	0,0042	d	7182	0,0241
d	9639	0,0377	č	3352	0,0130	d	9637	0,0330	e	24691	0,0829
ď	182	0,0007	d	8764	0,0340	e	20509	0,0702	ě	5201	0,0175
e	20371	0,0796	ď	601	0,0023	ę	4613	0,0158	dź	2668	0,0090
é	2032	0,0079	dz	212	0,0008	f	416	0,0014	f	276	0,0009
ě	4719	0,0184	dž	5	0,0000	g	4387	0,0150	g	607	0,0020
f	213	0,0008	e	20564	0,0798	h	3199	0,0110	h	5540	0,0186
g	541	0,0021	é	1276	0,0049	i	25264	0,0865	ch	3579	0,0120
h	4207	0,0164	f	366	0,0014	j	5964	0,0204	i	13527	0,0454
ch	2460	0,0096	g	581	0,0023	k	9499	0,0325	j	17213	0,0578
i	11365	0,0444	h	4034	0,0156	l	5354	0,0183	k	10640	0,0357
í	6301	0,0246	ch	2695	0,0105	ł	8623	0,0295	ł	4024	0,0135
j	5552	0,0217	i	15166	0,0588	m	8510	0,0291	l	7238	0,0243
k	10193	0,0398	í	2358	0,0091	n	14275	0,0489	m	9647	0,0324
l	15223	0,0595	j	4270	0,0166	ń	406	0,0014	n	16201	0,0544
m	8320	0,0325	k	10010	0,0388	o	22229	0,0761	ń	505	0,0017
n	14183	0,0554	l	13204	0,0512	ó	2052	0,0070	o	27097	0,0909
ň	253	0,0010	Í	30	0,0001	p	8933	0,0306	ó	2813	0,0094
o	20618	0,0806	ľ	1456	0,0056	r	13344	0,0457	p	8425	0,0283
ó	86	0,0003	m	8569	0,0332	s	12627	0,0432	q	0	0,0000
p	8252	0,0322	n	12842	0,0498	ś	1851	0,0063	r	10995	0,0369
q	0	0,0000	ň	642	0,0025	t	10120	0,0347	ř	2241	0,0075
r	8477	0,0331	o	23869	0,0926	u	6564	0,0225	s	12224	0,0410
ř	3290	0,0129	ó	100	0,0004	w	12876	0,0441	š	5625	0,0189
s	12174	0,0476	ô	320	0,0012	y	11170	0,0383	t	11500	0,0386
š	2650	0,0104	p	8293	0,0322	z	18622	0,0638	ć	4135	0,0139
t	12586	0,0492	q	0	0,0000	ź	254	0,0009	u	10113	0,0339
ť	169	0,0007	r	12233	0,0475	ż	2548	0,0087	v	0	0,0000
u	9147	0,0357	ŕ	94	0,0004				w	14719	0,0494
ú	188	0,0007	s	11959	0,0464				x	0	0,0000
ů	892	0,0035	š	2772	0,0108				y	7697	0,0258
v	11312	0,0442	t	11548	0,0448				z	7725	0,0259
w	7	0,0000	ť	1722	0,0067				ž	3412	0,0114
x	12	0,0000	u	7389	0,0287						
y	5219	0,0204	ú	1867	0,0072						
ý	2098	0,0082	v	12137	0,0471						
z	5338	0,0209	w	6	0,0000						
ž	2932	0,0115	x	10	0,0000						
			y	3697	0,0143						

			ý	2498	0,0097						
			z	5839	0,0226						
			ž	2419	0,0094						
	255880	1		257795	1			291979	1	297996	1

Glottometrics 18, 2009, 69-76

Nominal suffixes in German press texts

Emília Nemcová, Trnava

Abstract. In the article four German newspapers and the nominal derivations in them are scrutinized. We search for a possible model of rank-frequency sequences, its properties, and try to corroborate a more general linguistic hypothesis.

Keywords: Nominal suffixes, German, ranking, comparison

The aim of this article is threefold: (1) to investigate the difference between the newspapers comparing some properties of the ranked sequences of suffixes; (2) to test the hypothesis of quantitative linguistics that if there is a closed class of linguistic entities, their use is controlled in such a way that the ordering of frequencies in decreasing manner yields a very regular sequence; (3) to search for the form of this sequence and try to express it formally.

The data of this investigation were taken from Becker (1995), who analyzed the words in 112 articles of four newspapers written in German, namely from *Frankurter Allgemeine Zeitung* (Germany), *Neue Zürcher Zeitung* (Switzerland), *Die Presse* (Austria) and *Neues Deutschland* (former GDR), all concerning economy and published in the years 1984-1986. In the course of twenty years the forms of the four German "press dialects" changed and a new investigation of present day newspapers could show whether there is a convergence or a divergence (there is no GDR any more), what are the tendencies in nominal suffixation and how does German develop as a whole.

The class of derivation affixes in German press consisted of the following set: {*-ung, -e, -ion, -schaft, -t, -ie, -er, -ent, -nis, -enz, -zeug, -it, -keit, -heit, -el, -ik, -al, -ität, -tum, -anz, -ur, -ium, -ismus, -ement, -um, eur, -or, -ant, -är, -ei, -ial, -age, -ier, -ose, -at, -igkeit, -elle, -ar, -ist, -arier, -ling, -ül, -itis, -ler, -in, -chen, -lein, prefix+zero morpheme*}. Some examples for each suffix are as follows: {*Regierung, Höhe, Kommission, Wirtschaft, Sicht, Industrie, Verbraucher, Präsident, Hindernis, Konferenz, Fahrzeug, Defizit, Möglichkeit, Einheit, Klausel, Kritik, Signal, Liquidität, Wachstum, Instanz, Korrektur, Gremium, Optimismus, Management, Jubiläum, Exporteur, Senator, Lieferant, Aktionär, Partei, Potential, Passage, Passagier, Prognose, Senat, Arbeitslosigkeit, Novelle, Kommissar, Journalist, Subventionitis, Wissenschaftler, Politikerin, Fähnchen, Zünglein, Gesetz*}. Not all of them occur in all newspapers but all occur with different frequency, nevertheless, the ranks may be approximately equal.

The basic data are presented in Table 1. Here, the order of the affixes is qualityative. The affixes do not have the same status. Many grammarians would eliminate some of them as non-productive, other ones as foreign, especially if the word has been borrowed as a whole, confixes, affixoids, etc. Nevertheless, the table displays one of the aspects of forming nouns in these four newspapers and yields data which are appropriate for scrutinizing the above problems.

1. Equality

In data like those in Table 1, it is not possible to use the chi-square test for homogeneity because many frequencies are too small. On the other hand, pooling some classes is rather haphazard and can be manipulated in some direction. Thus the only possibility is to use the

ranks and test their homogeneity using a non-parametric test. The reordering of frequencies is presented in Table 2.

Table 1
Frequencies of affixes in four German newspapers (data from H. Becker 1995)

Suffix	FAZ	NZZ	PRESSE	ND
-ung	260	299	241	324
-e	78	106	49	40
-ion	42	71	44	41
-schaft	34	15	20	23
-t	34	26	30	12
-ie	21	14	18	26
-er	18	29	28	50
-ent	17	27	16	58
-nis	17	18	8	33
-enz	11	11	7	3
-zeug	11	4	3	9
-it	10	4	6	1
-keit	9	15	18	21
-heit	8	11	7	17
-el	8	14	4	5
-ik	8	13	5	28
pref.+0-morpheme	6	12	5	21
-al	6	6	8	-
-ität	5	8	12	12
-tum	5	5	-	-
-anz	4	-	2	1
-ur	4	-	-	3
-ium	3	5	2	5
-ismus	3	4	8	7
-ement	3	-	3	1
-um	2	5	4	1
-eur	2	-	3	1
-or	2	2	7	5
-ant	2	1	-	7
-är	2	1	2	5
-ei	1	1	-	8
-ial	1	3	1	2
-age	1	2	-	-
-ier	1	1	-	2
-ose	1	2	-	-
-at	1	8	3	32
-igkeit	1	1	3	5
-elle	1	-	1	-
-ar	-	4	1	-
-ist	-	2	2	4
-arier	-	2	-	-
-ling	-	1	-	2
-ül	-	1	-	-
-itis	-	1	-	-
-ler	-	-	1	1
-in	-	-	-	4
-chen	-	-	-	1
-lein	-	-	-	1

Table 2
The ranks of individual suffixes in four newspapers

Suffix	FAZ	NZZ	PRESSE	ND
-ung	1	1	1	1
-e	2	2	2	5
-ion	3	3	3	4
-schaft	4.5	8.5	6	9
-t	4.5	6	4	14.5
-er	7	4	5	3
-ie	6	10.5	40.5	8
-ent	8.5	5	7	2
-nis	8.5	7	10	6
-enz	10.5	14.5	13	27.5
-zeug	10.5	23.5	22	16
-it	12	23.5	15	35
-keit	13	8.5	40.5	11.5
-heit	15	14.5	13	13
-el	15	10.5	18.5	22
-ik	15	12	16.5	7
zero	17.5	13	16.5	11.5
-al	17.5	18	10	43.5
-ität	19.5	16.5	8	14.5
-tum	19.5	20	40.5	43.5
-anz	21.5	44	27	43.5
-ur	21.5	44	40.5	27.5
-ismus	24	23.5	10	18.5
-ium	24	20	27	22
-ement	24	44	22	35
-um	28	20	18.5	35
-eur	28	44	22	35
-or	28	29	13	22
-ant	28	35.5	40.5	18.5
-är	28	35.5	27	22
-ei	34.5	35.5	40.5	17
-ial	34.5	26	31	30
-age	34.5	29	40.5	43.5
-ier	34.5	35.5	40.5	30
-ose	34.5	29	40.5	43.5
-at	34.5	16.5	22	10
-elle	34.5	44	31	43.5
-igkeit	34.5	35.5	22	22
-ar	43.5	23.5	31	43.5
-ist	43.5	29	27	25.5
-arier	43.5	29	40.5	43.5
-ling	43.5	35.5	40.5	30
-ül	43.5	35.5	40.5	43.5
-itis	43.5	35.5	40.5	43.5
-ler	43.5	44	27	35
-in	43.5	44	40.5	25.5
-chen	43.5	44	40.5	35
-lein	43.5	44	40.5	35

Equal ranks are ties which must be taken into account. In order to compare the homogeneity of ranks we compute Kendall´s concordance coefficient (cf. &&&Gibbons 1971: 250-257; Bortz, Lienert, Boehnke 1990: 465-470). Inserting the empirical values computed from Table 2 into the formula

$$(1) \qquad W = \frac{12D}{m^2(N^3 - N) - m\sum_{j=1}^{m} V_j}$$

where D is the sum of squared deviations of ranks from their mean, $m = 4$, $N = 48$, and V is a well known function of the lengths of ties. We obtain here $W = 0.7652$ corroborating the fact that the use of suffixes is homogeneous in all the newspapers. The chi-square test yielding 143.86 with 47 degrees of freedom corroborates this fact even though the frequencies of suffixes in individual newspapers are not identical.

However, even if the ranking may be equal, the concentration of frequencies may be different. Without performing tests for difference, we compute the relative repeat rate and the relative entropy of frequencies and obtain the results in Table 3. The formulas used are as follows. The repeat rate is defined as

$$(2) \qquad R = \sum_{i=1}^{K} \left(\frac{f_i}{N} \right)^2 ,$$

where K is the size of the inventory of suffixes, f_i the individual absolute frequencies (i = 1,2,…,K). The relative entropy is given as

$$(3) \qquad H_{rel} = \frac{H}{ld\, K}$$

where H is defined as

$$(4) \qquad H = -\sum_{i=1}^{K} \frac{f_i}{N} ld \frac{f_i}{N} .$$

Table 3
Repeat rates and relative entropies in the newspapers

Newspaper	Inventory K	Repeat rate	Rel. Entropy
FAZ	38	0.1933	0.6636
NZZ	39	0.1929	0.6532
PRESSE	34	0.2012	0.6802
ND	38	0.1818	0.6893

The greater the repeat rate – the smaller the relative entropy – the more a newspaper concentrates on a small number of suffixes. Thus only the Austrian newspaper “Die Presse” seems to have a slight divergence (a greater concentration) because it uses a smaller number of suffixes. But by and large the use of nominal suffixes is homogeneous.

We can conclude that from this restricted point of view the language of the press in the four official “dialects” in the eighties of the past century was homogeneous.

2. Regularity

Are the noun-forming suffixes a monolithic class? As can be seen in Table 1, not all suffixes are present in all newspapers, hence the classes are different. Do they form monolithic classes in spite of this fact? A class is monolithic if the frequencies of the individual elements (classes) are used in such a way that the differences between frequencies can be captured by a closed formula which serves as a model. Here we can proceed in two ways: we consider the ranked frequencies of suffixes as simple sequences or we consider them as discrete distributions. The difference is not essential for modelling, the latter way simply reduces the number of models because distributions must be normalized and in case of small inventories, like in our case, they should be truncated on the right side. Simple functions have the advantage of no normalization and no truncation; the validity of the function is restricted to the given definition domain.

If we look at the frequencies in Table 1, we see that the first suffix alone covers at least 39% of frequencies in all newspapers. This leads to the assumption that the process of forming nominal derivates is controlled by a latent mechanism. In order to capture it, we can consider the suffixes as urns in which balls are thrown randomly. However, the urns have the property of becoming the more attractive, the more balls are already in them. This leads to a stochastic process resulting in the negative binomial distribution (the tendency to repel further balls the more balls there already are results in the binomial distribution, and the neutral behaviour of the urns results in the Poisson distribution). If our hypothesis based on analogy is reasonable, we must take into account two circumstances: the number of suffixes is finite and very small, hence the distribution must be truncated at the right side; further, the ranking begins with 1 – being a deliberate decision – hence we must displace the distribution one step to the right. As a result we obtain the 1-displaced right truncated negative binomial distribution defined as

$$P_x = \binom{k+x-2}{x-1}\frac{q^{x-1}}{S(R)}, \quad x = 1,2,...,R+1 \tag{5}$$

where k and q are parameters and $S(R)$ is the normalizing factor defined as $S(R) = \sum_{j=1}^{R+1}\binom{k+j-2}{j-1}q^{j-1} = {}_2F_1(k,-R;-R,q)$, where $F(.)$ is the hypergeometric function. Fitting this distribution to the above data by means of a software (Altmann-Fitter 1994) we obtain the results presented in Table 4. As can be seen, all fittings are highly significant, the ND data somewhat weaker than the other ones; nevertheless, the fitting is more than satisfactory. Needless to say, some other distributions had been adequate in all cases, too (e.g. negative hypergeometric, Zipf-Alekseev) but the generally expected ones like Zipf-Mandelbrot or Zipf (zeta) distributions are not adequate for the ND data. The ND data seem to escape from the common attractor and tend to the Hyperpascal distribution which is a generalization of the negative binomial (cf. Wimmer, Altmann 1999).

Table 4
Fitting the negative binomial distribution to press data

	FAZ		**NZZ**		**PRESSE**		**ND**	
Rank *i*	f_i	NP_i	f_i	NP_i	f_i	NP_i	f_i	NP_i
1	260	256.14	299	297.39	241	235.00	324	301.91
2	78	77.45	106	91.55	49	68.74	58	97.62
3	42	48.44	71	57.43	44	42.72	50	62.12
4	34	35.50	29	42.13	30	31.24	41	46.01
5	34	27.94	27	33.16	28	24.55	40	36.48
6	21	22.88	26	27.15	20	20.11	33	30.07
7	18	19.24	18	22.81	18	16.90	32	25.41
8	17	16.47	15	19.51	18	14.48	28	21.85
9	17	14.2870	15	16.91	16	12.57	26	19.04
10	11	12.52	14	14.81	12	11.03	23	16.75
11	11	11.07	14	13.07	8	9.76	21	14.85
12	10	9.85	13	11.62	8	8.69	21	13.26
13	9	8.81	12	10.38	8	7.78	17	11.89
14	8	7.91	11	9.31	7	7.00	12	10.72
15	8	7.14	11	8.39	7	6.33	12	9.70
16	8	6.46	8	7.59	7	5.74	9	8.80
17	6	5.87	8	6.88	6	5.22	8	8.01
18	6	5.34	6	6.25	5	4.76	7	7.31
19	5	4.87	5	5.70	5	4.35	7	6.69
20	5	4.46	5	5.20	4	3.98	5	6.13
21	4	4.08	5	4.76	4	3.65	5	5.63
22	4	3.75	4	4.36	3	3.36	5	5.17
23	3	3.44	4	4.00	3	3.09	5	4.77
24	3	3.17	4	3.68	3	2.85	5	4.39
25	3	2.92	4	3.38	3	2.63	4	4.06
26	2	2.69	3	3.12	3	2.43	4	3.75
27	2	2.49	2	2.87	2	2.25	3	3.47
28	2	2.30	2	2.65	2	2.08	3	3.21
29	2	2.13	2	2.45	2	1.93	2	2.98
30	2	1.97	2	2.26	2	1.79	2	2.76
31	1	1.83	2	2.10	2	1.66	2	2.57
32	1	1.69	1	1.94	1	1.55	1	2.39
33	1	1.57	1	1.80	1	1.44	1	2.22
34	1	1.46	1	1.67	1	1.34	1	2.06
35	1	1.36	1	1.55			1	1.92
36	1	1.26	1	1.44			1	1.79
37	1	1.17	1	1.33			1	1.67
38	1	1.09	1	1.24			1	1.56
39			1	1.15			1	
	k = .3189,q = .9483 X^2 = 5.49, DF = 34 P ~ 1.0		k = .3252, q = .9466 X^2 = 17.29, DF = 35 P = 0.99		k = .3078, q = .9505 X^2 = 10.10, DF = 30 P ~ 1		k = .3406, q = .9494 X^2 = 39.15, DF = 34 P = 0.25	

The (iteratively estimated) parameters of all data are almost identical. Using Ord´s criterion and plotting the <*I*,*S*> points in a Cartesian coordinate system, where $I = m_2/m'_1$, $S = m_3/m_2$ (first raw moment and second and third central moments) we can see in Figure 1 that the points lie in the domain of the negative hypergeometric distribution but in the near vicinity of the negative binomial line $S = 2I-1$ which begins at the point <1,1>. The individual coordinates are: FAZ = <8.77, 14.45>, NZZ = <9.06, 14.76>, PRESSE = <8.35, 13.00>, ND = <8.12, 12.62>.

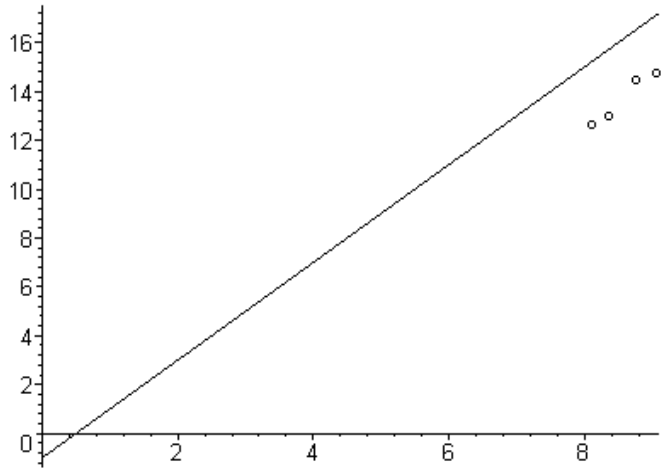


Figure 1. Ord´s scheme for nominal suffixes in German press texts

Since Zipf´s and Mandelbrot´s approach fail in two cases (Zipf with ND data and Zipf-Mandelbrot in PRESSE and ND data) we may ask whether the simple power function representing Zipf´s law could be adequate. As a matter of fact, in all cases we obtain excellent results with very high determination coefficients D:

FAZ	$y = 255.9865x^{-1.4775}$,	D = 0.99
NZZ	$y = 297.7476x^{-1.4208}$,	D = 0.996
PRESSE:	$y = 236.0361x^{-1.6426}$,	D = 0.98
ND:	$y = 313.2356x^{-1.5756}$,	D = 0.96

Here, *x* is the rank, *y* is the frequency and the fit is very good. This fact shows that treating rank-frequency data one can restrict oneself to simple functions, as has been done by Zipf himself. Almost identical results can be attained using the Popescu approach (Popescu, Altmann, Köhler 2009) with only one exponential component.

3. Summary

The questions asked above can be answered as follows: the use of nominal suffixes is not significantly different in the four local press texts; there are relatively clear tendencies and preferences in nominal derivation in German. The model of the rank-frequency sequence can be substantiated by a stochastic process resulting in the negative binomial distribution. Though other distributions may be used, too, we preliminarily restrict ourselves to this model. Considering the rank-frequency sequence as a simple function, the original Zipf´s approach yields excellent results. The hypothesis concerning the control mechanism of rank-frequency formation for any classes can be considered as locally corroborated.

References

Becker, H. (1995). *Die Wirtschaft in der deutschsprachigen Presse*. Frankfurt: Lang.

Bortz, J., Lienert, G.A., Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin-Heidelberg-New York: Springer.

Gibbons, J.D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

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

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

Software

Altmann-Fitter (1994). *Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen.* Lüdenscheid: RAM-Verlag..

Glottometrics 18, 2009, 77-96

History of Quantitative Linguistics

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

XXXVI. Quantitative Hypothesen bei Mikołaj Kruszewski

Der biographische Hintergrund des bekannten polnischen Sprachwissenschaftlers und prominenten Vertreters der Kazaner sprachwissenschaftlichen Schule Mikołaj Kruszewski (1851-1887) ist in einer Reihe von Arbeiten ausführlich behandelt (vgl. Berezin 1968, Jakobson 1971, Koerner 1986 bzw. 1989, Radwańska-Williams 1993, Jarceva 1998, Alpatov 2005: 114ff u.a.). Daher können wir uns an dieser Stelle auf die wichtigsten Eckpunkte konzentrieren. Geboren am 6. (18.) Dezember 1851 in Łuck (heutige Ukraine) studierte Kruszewski an der Warschauer Universität Philosophie, Logik, Psychologie. Mit Sprachwissenschaft beschäftigte er sich nur in zweiter Linie. Nach Abschluss seines Studiums im Jahr 1875 trat er eine Stelle als Lehrer für klassische Sprachen in Troick (Orenburg, Russland) an. Im gleichen Jahr kam er mit Jan Baudouin de Courtenay (1845-1929) in Kazan' in Kontakt. In weiterer Folge bildete Mikołaj Kruszewski nach seiner Übersiedelung nach Kazan' (1878) mit Baudouin de Courtenay das personelle Grundgerüst der Kazaner Schule[1]. In diese Zeit, bis zu seinem frühen Tod im Jahr 1887, fällt auch die produktivste wissenschaftliche Periode von Kruszewski.

Seine wichtigsten monographischen Beiträge (ausführliches Schriftenverzeichnis siehe Kruševskij 1998: 272-275) sind „Über die Lautabwechslung"[2] (1881) und die theoretische linguistische Schrift "Očerk nauki o jazyke // Abriß der Wissenschaft von der Sprache[3]" aus dem Jahr 1883. Aus theoretischer und inhaltlicher Sicht wird Kruszewski heute mit folgenden sprachwissenschaftlichen Bereichen verbunden: Hervorgehoben wird seine Vorreiterrolle für die strukturalistische Sprachwissenschaft, die sich aus der frühen Diskussion von Konzepten wie „Statik" vs. „Dynamik", „Syntagmatik" vs. „Paradigmatik", der Untersuchung von lebenden Sprachen, der Zeichenhaftigkeit der Sprache usw. ablesen lässt. Wichtig sind auch Kruszewskis Überlegungen zur Morphophonologie. Er unterschied phonetische Alternationen von historisch auftretenden Lautwandel und führte darüber hinaus bereits formale Regeln – ähnlich dem generativen Ansatz – für diese Phänomene ein (vgl. dazu Klausenburger 1978).

Trotz einer gerechtfertigten historiographischen Integration von Kruszewskis theoretischen Ansätzen in strukturalistische und generative Schulen, darf nicht übersehen werden, dass er ebenso dem junggrammatischen Paradigma verpflichtet bleibt. Dies lässt sich an seiner Vorliebe für sprachliche Gesetzmäßigkeiten, die er in das Zentrum seines linguistischen Interesses stellt, ablesen. Es geht ihm dabei z.B. um „statičeskij zakon zvuka // stati-

[1] Unklar bleibt der Grad der gegenseitigen wissenschaftlichen Beeinflussung zwischen Baudouin de Courtenay und Mikołaj Kruszewski (vgl. Koerner 1995: xiv, vgl. auch die Eigeneinschätzung in Courtenay 1888ff.). Zuweilen wird M. Kruszewski als Schüler von Baudouin de Courtenay angesehen. Dennoch lassen sich bei ihm eine Reihe von eigenständigen Ansätzen erkennen, die es erlauben von einem eigenen wissenschaftlichen Profil zu sprechen.

[2] Die englische Übersetzung dieser beiden Monographien findet sich in Kruszewski (1995).

[3] Auszüge aus dem „Očerk nauki […]" sind auf Deutsch unter dem Titel „Prinzipien der Sprachentwicklung" in der „Internationalen Zeitschrift für allgemeine Sprachwissenschaft" erschienen.

sches Lautgesetz“, um „statičeskij zakon zvukovogo sočetanija // statisches Gesetz der Lautverbindung“ und um „dinamičeskie zakony zvuka // dynamische Lautgesetze“ usw.

Es lassen sich aber in den Arbeiten von Kruszewski eine ganze Reihe von Andeutungen und Hypothesen über die quantitative Struktur von Sprachen und die Diskussion von quantitativen „Gesetzmäßigkeiten“ nachweisen. Diese Anregungen sind aus heutiger Sicht durchaus attraktiv und können als Anstoß für die Ausformulierung von neuen Hypothesen nützlich sein. Sie verdienen es im gegebenen Zusammenhang näher vorgestellt zu werden. Wichtigste Quelle dafür ist sein zentrales theoretisches Werkes „Očerki nauki o jazyke // An outline of linguistic science“ (1883); im Folgenden wird diese Arbeit nach den gesammelten Werken in Kruševskij (1998) zitiert[4].

Die epistemologische Position von Kruszewski ist widersprüchlich: Einerseits spricht er sich für eine deduktive, theoriengeleitete Sprachwissenschaft aus, anderseits aber für einen empirisch-induktiven Ansatz (vgl. Kruševskij 1998: 98). Auch wenn Kruszewskis Überlegungen in diesem Fall von einer Synthese (im Sinne des Einschiebens einer Zwischenebene in Form der Abduktion) weit entfernt ist, sind seine Arbeiten genau von dieser Position geprägt: entweder sind seine Überlegungen rein theoretischer Natur, oder aber empirische Generalisierungen, die in der Regel anhand einer Fülle von empirischen Beobachtungen begleitet sind.

Eine hervorragende Rolle schreibt Kruszewski – im Sinne von empirischen Generalisierungen – linguistischen Gesetzmäßigkeiten zu. So kommt er im Zusammenhang um die Diskussion statischer Lautgesetze (Kruševskij 1998: 107ff.), die auf eine phonologischen Interpretation von Lauten (Gleichförmigkeit und Wiederholbarkeit des Lautstromes) hinauslaufen, auf ein „statičeskij zakon zvukovogo sočetanija // statisches Gesetz von Lautverbindungen“ zu sprechen. Dahinter verbirgt sich in heutiger Terminologie nichts anderes als die von Sprachen eingeschränkte Ausnützung der Bildung von Phonem- bzw. Lautverbindungen. D.h., angesprochen werden phonotaktische/distributionelle Probleme, die in Summe auch innerhalb der quantitativen Linguistik bislang nur in Teilbereichen bearbeitet worden sind. Vgl. u.a. Altmann/Lehfeldt (1980: 217ff.), Kleinlogel/Lehfeldt (1972) und den heutigen Erkenntnisstand zu distributional gaps, der Anzahl von Phonemverbindungen in Relation zum Phoneminventar (vgl. dazu Strauss/Fan/Altmann 2008: 5).

Interessant sind Kruszewskis Andeutungen zu eingeschränkten Kombinationsmöglichkeiten[5] von Phonemen/Lauten vor allem auch deshalb, weil damit der Versuch einer Begründung einhergeht: Es ist dies für ihn das unbewusste Streben des Sprechers nach Sprachökonomie und der Verminderung des physiologischen Aufwandes (vgl. Kruševskij 1998: 111). Ein Ansatz der Sprachökonomie also, der im Übrigen ebenso bei Hermann Paul, Hugo Schuchardt und vielen anderen Linguisten am Ende des 19. Jahrhunderts nachweisbar ist. Auch heute, wenn auch in etwas elaborierter Form, sind sprachökonomische Faktoren ein theoretischer Grundbestandteil der synergetischen und quantitativen Linguistik.

Kruszewski war sich nicht der erklärenden Kraft der Sprachökonomie bewusst, sondern auch der Rolle der Häufigkeit von linguistischen Einheiten im Allgemeinen. Dazu lassen

[4] Für eine Analyse aller weiteren Primärschriften von Kruszewski vgl. Adamska-Sałaciak (2005).

[5] Kruševskij geht auch detaillierter auf die diachrone Herausbildung von Lautverbindungen und Assimilationen ein. Darüber hinaus verweist er auf die – in der Regel akustisch begründbare – Unmöglichkeit bzw. absolute Absenz von Lautverbindungen in einigen Sprachen und das Vorhandensein bestimmter Lautverbindungen nur in Lehnwörtern bzw. an Morphemfugen. Dieses theoretische Gerüst der Untersuchung von phonotaktischen Problemen ist ebenfalls in einer der ersten russischen sprachstatistischen Monographie von Čistjakov/ Kramarenko (1929) analysiert worden (vgl. dazu Kelih 2008: 69-74). Dass Kruszewski den entsprechenden theoretischen Input geliefert hat, lässt sich zum einen an der ähnlichen Terminologie und Forschungsfrage ablesen. Zum anderen zeigt sich dies an den Versuchen einer quantitativen Analyse der kanonischen Struktur von Wortformen und Morphemen. Vgl. dazu Kruševskij (1998: 157ff) und Čistjakov/Kramarenko (1929: 42). In beiden Fällen geht es um Anfangs- und Endgruppen von Lautverbinungen in Wortformen.

sich bei ihm eine Reihe von Andeutungen finden. Erstens hat seiner Ansicht (vgl. Kruševskij 1998: 169) nach die unterschiedliche Ausprägung der Häufigkeit auf syntagmatischer Ebene die Funktion die „Speicher“- und „Reproduktionsfähigkeit“ des Sprechers zu erleichtern. Demnach werden häufig vorkommende sprachliche Elemente seitens des Sprechers ökonomischer „re-produziert“, als weniger oft vorkommende.

Zweitens verweist Kruszewski (1998: 169) auf den Suppletivismus sprachlicher Formen, welchen er wiederum mit der Häufigkeit in Verbindung bringt: Seiner Meinung nach ist dieses Phänomen vor allem bei frequenten Wortformen zu beobachten. Auch wenn dieses Phänomen ein doch recht bekanntes Phänomen[6] ist, so muss doch festgehalten werden, dass es bislang nur wenige systematische empirische Untersuchungen dazu gibt. Vgl. u.a. die Arbeit von Corbett et al. (2001) zu Wechselbeziehungen zwischen Irregularitäten in der russischen Morphologie und der Häufigkeit, die sich aus methodologischer Sicht wohltuend von anderen Arbeiten aus dem Kreis des „American frequentism“ (vgl. u.a. Bybee/Hooper 2001) abhebt.

Drittens ist auf den Zusammenhang zwischen der „Bedeutungsmenge“ und der Häufigkeit von Wörtern zu verweisen, der an dieser Stelle resümierend dargestellt werden kann (ausführlich dazu Kelih 2008b). Es geht um semantische Fragen der Bedeutungserweiterung und –verengung, die auf diachroner bzw. synchroner Ebene zu beobachten sind. Das von Kruševskij (1998: 206) in diesem Zusammenhang formulierte „Zakon obratnogo otnošenija meždu ob"emom i soderžaniem // Gesetz der reziproken Relation zwischen dem Umfang und dem Inhalt“ eines Wortes beinhaltet Aussagen über die Verwendungshäufigkeit/Frequenz und semantische Eigenschaften eines Wortes. Denn, so Kruszewski „je häufiger ein Wort verwendet wird, desto weniger Inhalt umfasst es“. Während hinsichtlich der Verwendungshäufigkeit/Frequenz von Wortformen die Aussage klar ist, ist das „weniger Inhalt“ in der Sekundärliteratur bislang unterschiedlich interpretiert worden. So interpretieren Levickij (2006) und zuvor Kuryłowicz (1962: 20) diese Passage als eine Beobachtung zum diachronen Verlauf des Verlusts bzw. des Erhalts der semantischen und morphologischen „Motivierung / motivirovannost’“ eines Wortes. Demgegenüber wird in Kelih (2008b) dafür plädiert, das „weniger Inhalt“ – unter Ausblendung der Diachronie – durchaus im Sinne von weniger „značenie/ Bedeutungen“ zu verstehen ist. Ansonsten ist diese Passage in dieser vagen Form kaum einer empirischen Überprüfung zuzuführen. Demnach würde die Häufigkeit eines Wortes mit der Anzahl von verschiedenen Bedeutungen (Polysemie) korrelieren. Damit ist man mit dem in der quantitativen Linguistik bekannten Problem der Wechselbeziehung zwischen der Häufigkeit eines Wortes auf syntagmatischer Ebene und seiner Polysemie auf paradigmatischer Ebene konfrontiert, welches in Tuldava (1979) bzw. Levickij (2005) ausführlich dargestellt ist.

Damit kann abschließend festgestellt werden, dass die Überlegungen und Andeutungen von Kruszewski keine Gesetzmäßigkeiten im Sinne der heutigen quantitativen Linguistik darstellen. Es fehlt ihnen vor allem eine konsistente Ausformulierung einer Hypothese und eine empirische Überprüfung der gemachten Aussagen mit der Hilfe von statistischen Methoden. Demnach sind Kruszewskis obige Andeutungen als geglückte linguistische Beobachtungen zu qualifizieren. Gleichzeitig ist sein Werk aber ein repräsentatives Beispiel dafür, dass bis in das 19. Jahrhundert zurückreichende Beobachtungen zur Laut- Wortstruktur und Bedeutungsmenge durchaus auch heute noch – nach einer entsprechenden Reformulierung – einen sinnvollen Ausgangspunkt von quantitativen Untersuchung darstellen können.

[6] Ähnliches hatte Schuchardt (1885: 25 bzw. 28) für den Zusammenhang zwischen der „Archaizität“ und der Häufigkeit sprachlicher Formen festgestellt.

Literatur

Altmann, G., Lehfeldt, W. (1980). *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer. [= Quantitative Linguisitcs, 7]

Adamska-Sałaciak, A. (2005). *Language change in the works of Kruszewski, Baudouin de Courtenay and Rozwadowski*. Poznań: Motivex.

Alpatov, V.M. (2005). *Istorija lingvističeskich učenij. Učebnoe posobie. 4-e izdanie, ispravlennoe i dopolnennoe*. Moskva: Jazyki Slavjanskoj Kul'tury.

Baudouin de Courtenay, I. (1888f.). Mikołaj Kruszewski, jego życie i prace naukove. In: *Prace filologiczne*, 2, 838-849. 3 (1889), 116-175. [zitiert nach Boduėn de Kurtenė, I.A. (1963): *Izbrannye trudy po obščemu jazykoznaniju. Tom 1*. Moskva: Akademija Nauk. u.d.T.: Nikolaj Kruševskij: ego žizn' i naučnye trudy, 146-202.]

Berezin, F.M. (1968). *Očerki po istorii jazykoznanija v Rossii (konec XIX načalo XX v.)*. Moskva: Nauka.

Bybee, J., Hooper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: John Benjamins Publishing Company. [= Typological Studies in Language, 45]

Corbett, G., Hippisley, A., Brown, D., Marriot, P. (2001). Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In: Bybee, J.; Hooper, P. (eds.) (2001), 199-226.

Čistjakov, V.F., Kramarenko, B.K. (1929). *Opyt priloženija statističeskogo metoda k jazykoznaniju. Vyp. I*. Krasnodar.

Jakobson, R.O. (1971). Značenie Kruševskogo v razvitii nauki o jazyke. In: Roman, J. (1971): *Selected writings. Vol. II: Word and Language*. The Hague: Mouton, 429-450.

Jarceva, V.N. (1998). N.V. Kruševskij - provozvestnik lingvistiki XX veka. In: Kruševskij, M. (1998): *Izbrannye raboty po jazykoznaniju*. Moskva: Nasledie, 4-24.

Kelih, E. (2008a). *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*. Hamburg: Kovač. [in Druck]

Kelih, E. (2008b). Semantische Gesetze: Der Fall M. Kruszewski. In: Altmann, G.; Zadorozhna, I.; Matskulyak, J. (eds.): *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th anniversary of Professor V.V. Levickij*. Chernivtsi: Knichi XXI, 226-232.

Klausenburger, J. (1978). Mikołaj Kruszewski's Theory of Morphophonology, in: *Historiographia Linguistica*, 5, 109-120.

Kleinlogel, A., Lehfeldt, W. (1972). Zur Problematik einer syntagmatisch-phonologischen Sprachklassifikation. In: Jäger, Siegfried (eds): *Linguistik und Statistik*. Vieweg: Braunschweig, S. 51-64. [= Schriften zur Linguistik, 6]

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin u.a.: Walter de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

Koerner, E.F.K. (1986): Kruszewski's Contribution to General Linguistic Theory. In: Kastovsky, D.; Szwedek, A. (eds.): *Linguistics across Historical and Geographical boundaries. ed. by, vol. I: Linguistic Theory and Historical Linguistic*. Berlin: de Gruyter, 53-57.

Koerner, K. (1995). Introduction. Mikołaj Kruszewski's contribution to general linguistics. In: Kruszewski, M. (1995), xi-xl.

Kruszewski, M. (1995). *Writings in general linguistics. Edited and with an introdcution by Konrad Koerner*. Amsterdam [u.a.]: Benjamins. [= Amsterdam studies in the theory and history of linguistic science, 1. Amsterdam classics in linguistics; 11]

Kruševskij, N. (1998). *Izbrannye raboty po jazykoznaniju*. Moskva: Nasledie.

Kuryłowicz, J. (1960). *Esquisses linguistiques.* Wrocław – Kraków. [russ. Übersetzung in: Kurilovič, E. (1962): *Očerki po lingvistike.* Moskva: Izdatel'stvo inostranoj literatury]

Levickij, V. (2005). Polysemie. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005), 458-464.

Levickij, V.V. (2006). *Semasiologija.* Vinnica: Nova Knyha.

Radwańska-Williams, J. (1993). *A paradigm lost: the linguistic theory of Mikołaj Kruszewski*. Amsterdam [u.a.]: Benjamins. [= Amsterdam studies in the theory and history of linguistic science: Series 3, Studies in the history of the language sciences; 72]

Schuchardt, H. (1885). *Über die Lautgesetze. Gegen die Junggrammatiker*. Berlin: Verlag Robert Oppenheim.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag. [= Studies in Quantitative Linguistics, 1]

Tuldava, Ju. (1979). O nekotorych kvantitativno-sistemnych charakteristikach polisemii. In: *Linguistica* XI, 107-141. [= Učenye zapiski Tartuskogo gosudarstvennogo universiteta, 502]

Emmerich Kelih, Graz

XXXVII. Friedrich Wilhelm Kaeding (1843-1928)

Geb. 18.9.1843 in Rathenow; gest. 29.8.1928. Gymnasium bis zur Sekunda, Militärdienst; 1858 Kreisgerichtsdeputation in Rathenow; 1968 nach Berlin; 1873 bei Reichsbank, dort 1882 Kalkulator, 1895 Oberkalkulator, 1899 Rechnungsrat, 1910 Geheimer Rechnungsrat. 1874 zusammen mit Dreinhöfer Gründung des Verbandes Stolzescher Stenographenvereine, 1. Vorsitzender. Aktiv an der Entwicklung des Systems Stolze-Schrey beteiligt und um Vermittlung zwischen den verschiedenen Stenographie-Schulen bemüht, nacheinander Mitglied verschiedener Stenographenorganisationen.

Kaeding ist für die Quantitative Linguistik deshalb sehr wichtig, weil auf seine Initiative das bedeutende und wegweisende *Häufigkeitswörterbuch der deutschen Sprache* (Kaeding [Hrsg.] 1897/98) zustande gekommen ist. Er selbst hat im September 1891 auf dem „Stolzetag“ in Berlin – einer Veranstaltung von Stenographen – beantragt, ausgedehnte Häufigkeitsuntersuchungen durchzuführen. Dem Antrag wurde zugestimmt; kurz danach übernahm der „Internationale Stenographentag diesen Antrag und beauftragte Kaeding mit seiner Ausführung. Aufgrund von Voruntersuchungen sollte ein Umfang von 20000000 Silben bzw. „fast 11 Millionen Wörter(n)“ (Kaeding 1897/98: 6) erreicht werden. Thema waren „Untersuchungen zur Feststellung der Häufigkeit der Wörter, Silben und Laute in der deutschen Sprache“ (Kaeding 1897/98: 7). Kaeding beklagt, dass mögliche Interessen der Sprachforscher mangels entsprechender Kooperationsbereitschaft der Linguisten weniger als denkbar befriedigt werden konnten. Die Arbeit wurde daher primär von den Interessen der Stenographen bestimmt, allerdings auch mit heftiger Kritik begleitet. Nicht aufgenommen wurden Eigennamen und Zahlzeichen, wohl aber ausgeschriebene Zahlwörter und Fremdwörter. Bei der Auswahl der zu zählenden Texte (nur laufender Text) wurde auf eine breite thematische Streuung über alle Wissensgebiete Wert gelegt. Die Ergebnisse wurden auszugsweise publiziert; „die Urschrift des ganzen Werkes geht nach beendeter Drucklegung der Auszüge in das Eigentum der Königlichen Bibliothek über“ (Kaeding 1897/98: 31). Ortmann (1978: 13) fand diese Unterlagen in der deutschen Staatsbibliothek in Ostberlin vor. Einen Eindruck von dem Aufwand, den das Unternehmen erforderte, erhält man durch die Zahl der eingesetzten Mitarbeiter: Allein für die Ersterfassung der fast 11 Millionen Wörter, bei der jedes

Wort auf einen eigenen Zettel geschrieben wurde, wurden 665 Personen eingesetzt; noch einmal etwa die gleiche Zahl befasste sich mit den Auswertungen.

Eine Beschreibung des Verfahrens und der Ergebnisse geben außer Kaeding selbst auch Aichele (2005: 16f.), Meier (1967: 8f.), Njock (1973: 23-28), und Ortmann (1975a: 5-7); eine genaue Auflistung der publizierten Ergebnisse gibt Ortmann (1975a: 23-26)[7]. Kaedings Werk wurde wieder aufgegriffen und fortgeführt von Helmut Meier (1964/ 1967). Auf der Grundlage von Kaedings Werk ist Morgan (1928) entstanden; es war Vorbild für ähnliche Unternehmen bei anderen Sprachen, so z.B. zum Japanischen (Ito 2005: 84). Njock (1973: 28-40) führt eine Reihe weiterer Werke in der Nachfolge Kaedings an. Zipf (1935/68: 23) stützt sich auf Kaeding, um auf den Zusammenhang zwischen Wortlänge und -häufigkeit hinzuweisen.

Zur Bedeutung für die Linguistik meint Aichele (2005: 16f.): „Auch wenn sich Kaeding von seiner Arbeit als Nebeneffekt Antworten auf offene Fragen der Sprachforschung erhoffte, so ist sein Werk trotz entsprechender Ankündigungen in einschlägigen Fachorganen in sprachwissenschaftlichen Kreisen seinerzeit auf keine große Resonanz gestoßen und später erst wieder von Helmut Meier (1964, 1967) aufgegriffen worden.“ Man muss ergänzen: Auch Helmut Meier war von Hause aus kein Sprachwissenschaftler, sondern vor allem Lehrer und zeitweise Dozent in Braunschweig (Best 2007). Aber nicht nur Meier hat Kaeding fortgeführt; man muss hier auch auf Ortmann (1975-1981) verweisen, der die knapp 8000 häufigsten Wörter aus Kaedings Häufigkeitswörterbuch auf vielfältige Weise in mehreren Bänden ausgewertet hat. Weitere Untersuchungen Ortmanns berücksichtigen ebenfalls Kaeding, ohne dass dieser wie in den genannten Werken dominiert. Schubenz (1979) wiederum analysiert Kaedings Material in Meiers Bearbeitung von (1964) für lernpsychologische Zwecke. Die Wortlängenverteilung des gesamten Datenmaterials von Kaeding (nach der Zahl der Silben pro Wort) folgt der Hyperpoisson-Verteilung (Best 2006: 41) und kann damit als eine weitere Stütze für die Hypothese herangezogen werden, dass Wortlängen wie andere Sprachphänomene Gesetzen unterliegen (Altmann 1985; Wimmer u.a. 1994).

Diese Andeutungen mögen genügen. In der Quantitativen Linguistik ist Kaeding ein Begriff, unter Stenographen ohnehin. Er ist ein weiteres Beispiel dafür, dass die Quantitative Linguistik immer wieder entscheidend von Leuten profitiert, die gar nicht aus der Linguistik kommen und dennoch für wesentliche Fortschritte unserer Disziplin, oft aber auch darüber hinaus noch für die in weiteren Wissenschaften sorgen.

Als ein Beispiel dafür, dass Kaedings Erhebungen noch weiter für die Quantitative Linguistik genutzt werden können, sei seine Darstellung der Häufigkeit von Interpunktionszeichen angeführt, allerdings beschränkt ausschließlich auf die Satzzeichen. An die Datei der Satzzeichen für alle Textgruppen zusammen kann man Altmanns Modell für beliebige Rangordnungen (Altmann 1993: 62)

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

anpassen, wie die folgende Tabelle zeigt.

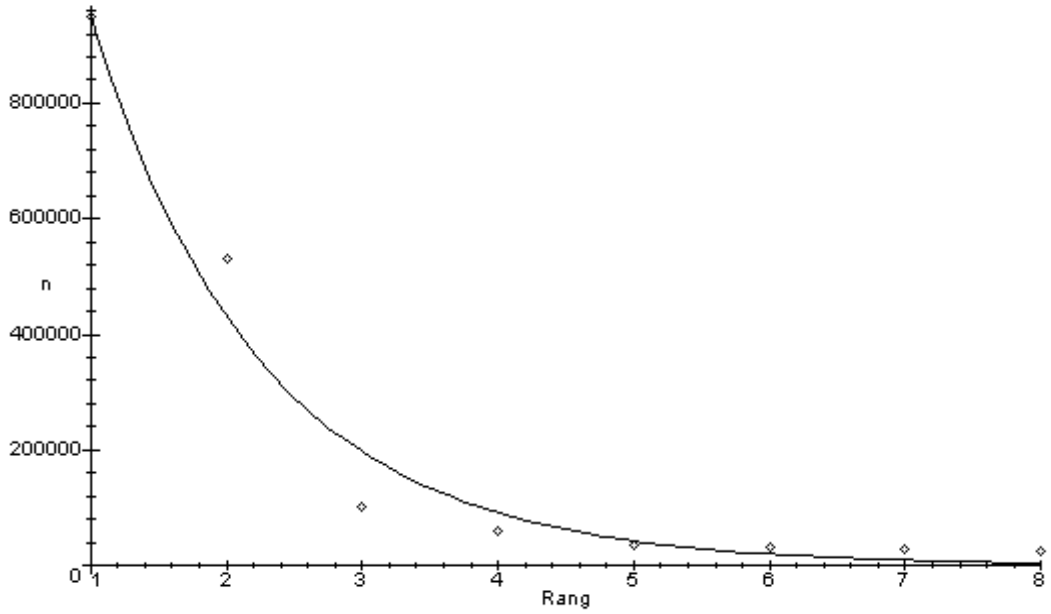
[7] Es wird darauf verzichtet, diese seitenlange Liste hier zu wiederholen.

Tabelle
Anpassung von Altmanns Modell für beliebige Rangordnungen
an Kaedings Gesamtdatei der Interpunktionszeichen (nur Satzzeichen)

Rang	Interpunktions-zeichen	Häufigkeit beobachtet	Häufikeit berechnet
1	Komma	948945	948945.00
2	Punkt	532512	430488.74
3	Anführungszeichen	101716	197116.44
4	Semikolon	59484	91080.54
5	Ausrufezeichen	35730	42459.53
6	Doppelpunkt	33325	19965.50
7	Apostroph	28899	9467.88
8	Fragezeichen	26414	4526.97
a = 125.85	b = 56.00	c = 948945.00	D = 0.97

Erläuterung zu der Tabelle:
a, *b*, *c*: Parameter des Modells. *D*: Determinationskoeffizient. Der Determinationskoeffizient ist akzeptabel, wenn $D \geq 0.80$, und gut mit $D \geq 0.90$; es handelt sich also mit $D = 0.97$ um ein sehr gutes Ergebnis.

Die Graphik veranschaulicht das gute Ergebnis noch einmal:



Graphik: Häufigkeit der Interpunktionszeichen bei Kaeding (Gesamtdatei)

Literatur

(Anmerkung: Nicht jeder der angegebenen Titel konnte am Original überprüft werden; etliche Angaben sind der zitierten Literatur entnommen. Stenographische Spezialliteratur ist in den wissenschaftlichen Bibliotheken nur wenig vertreten. Man findet sie schwerpunktmäßig in

speziellen Sammlungen zur Stenographie in folgenden Bibliotheken: *Bibliothek des Niedersächsischen Landtages* in Hannover, *Forschungs- und Ausbildungsstätte für Kurzschrift und Textverarbeitung in Bayreuth e.V.* sowie in der Stenographischen Sammlung der *Sächsischen Landes- und Universitätsbibliothek Dresden*; vieles auch in der *Thüringer Universitäts- und Landesbibliothek Jena*.

Die Angaben zu Kaedings Werken sind nicht vollständig und müssen das für die Zwecke der Quantitativen Linguistik auch nicht sein; sie sollen lediglich einen Eindruck von seinem Schaffen vermitteln.)

Zu Kaedings Häufigkeitswörterbuch

Amsel, Georg (1896). Über Kaedings Häufigkeitsuntersuchungen. In: Johnen, Christian (Hrsg.), *Festbuch zur hundertjährigen Jubelfeier der deutschen Kurzschrift : [Gewidm. dem Andenken ihrer Begründer Friedrich Mosengeil u. Karl Gottlieb Horstig] / Zur Mosengeilfeier auf dem 4. Verbandstage für vereinfachte deutsche Stenographie (System Schrey) zu Bonn am 28. Juni 1896* (S. 157-164). Berlin: Verlag von Ferdinand Schrey.

Amsel, Georg, & Kaeding, Friedrich Wilhelm (1896). Zur Statistik des deutschen Wortschatzes. *Zeitschrift des königlich preußischen statistischen Bureaus 36, 239-264.* (Entspricht nach Ortmann 1978, XL, Anmerkung 26, weitgehend der Einleitung zu Kaedings Häufigkeitswörterbuch.)

Frangen, Werner (2000). Lauthäufigkeiten nach Kaeding. http://www.forschungsstaette.de/PDF/Kaeding.pdf.

Gutzmann, Hermann (1898). Über das Häufigkeitswörterbuch der deutschen Sprache und seine Wichtigkeit für das Ablesen der Schwerhörigen und Ertaubten. *Medizinisch-pädagogische Monatsschrift für die gesamte Sprachheilkunde 8. Jg., Juni-Heft, 161-165.*

Heyne, M. (1900). Besprechung zu Kaeding, Häufigkeitswörterbuch. *Zeitschrift für deutsches Altertum und deutsche Literatur. Bd. N.F. 32 = 44/ Beilage: Anzeiger für deutsche Sprache und deutsche Litteratur/ Bd. XXVI, 1, 78-79.*

Kaeding, Friedrich Wilhelm (1892). Ueber Frequenzuntersuchungen. *Magazin für Stenographie XIII, 9-14.*

Kaeding, Friedrich Wilhelm (1892). Über Häufigkeitsuntersuchungen (Häufigkeit der Buchstaben, Wortstämme, Vor- und Nachsilben, Wortverbindungen und Laute der deutschen Sprache, Notwendigkeit dieser Untersuchungen und Art der Ausführung aller dazu gehörenden Arbeiten). *Magazin für Stenographie XIII, 177-182, 195-201.*

Kaeding, Friedrich Wilhelm (1893/4). Ueber die Einrichtung und den gegenwärtigen Stand der Häufigkeitsuntersuchungen. *Magazin für Stenographie XIV, 364-369, 381-385; XV/ 1894, 95-98, 113-117, 126-131.*

Kaeding, Friedrich Wilhelm (1895). Über die Häufigkeitsuntersuchungen der deutschen Sprache. *Magazin für Stenographie XVI, 71-76, 83-88, 103-108, 132-137, 153-155, 186-189, 201-203, 216-220, 234-237.*

Kaeding, Friedrich Wilhelm [Hrsg.] (1897/98). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Beiheft zu Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.* (Anmerkung: Die Titel von Band 1 und Band 2 enthalten nur die Jahreszahl 1897; der Gesamttitel das Jahr 1898.)

Kaeding, Friedrich Wilhelm (1898). Über die zweckmäßigste Zählmethode bei der wissenschaftlichen Kritik stenographischer Systeme sowie über den Wert der Schulschriftkürzungen im Einigungssystem Stolze-Schrey: Vortrag.

Kaeding, Friedrich Wilhelm (1899). Das Häufigkeitswörterbuch und die Geläufigkeitsuntersuchungen. *Magazin für Stenographie XX, 83-87, 90-94, 115-119, 129-133, 153-158.*

Meier, Helmut (1964, ²1967). *Deutsche Sprachstatistik.* Zweite erweiterte und verbesserte Auflage. Hildesheim: Olms.

Meissner-Luckenwalde (1905). F.W. Kaedings Häufigkeitswörterbuch der deutschen Sprache für die Schule verwertet. *Der Deutsche Stenograph V, 415-417.* (Kein Vorname genannt, auch nicht als Kürzel.)

Morgan, B. Q. (1928). *German Frequency Word Book, based on KAEDING'S Häufigkeitswörterbuch der deutschen Sprache.* New York: Macmillan.

Njock, Pierre Emmanuel (1973). *La lexicométrie allemande: 1898-1970.* Quebec: Centre international de recherches sur le biliguisme, Publication B-37.

Ortmann, Wolf Dieter (1975a). *Hochfrequente deutsche Wortformen I. 7995 Wortformen der KAEDING-Zählung, rechnersortiert in alphabetischer und rückläufiger Folge, nach Häufigkeit und Hauptwortarten.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1975b). *Beispielwörter für deutsche Ausspracheübungen. 7952 hochfrequente Wortformen der KAEDING-Zählung, rechnersortiert nach Einzellauten, Lautverbindungen, Silbenzahl und Akzentposition.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1975b). *Beispielwörter für deutsche Leseübungen. 7995 hochfrequente Wortformen der KAEDING-Zählung, rechnersortiert nach Graphem-Phonem--Beziehungen.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1975d). *Beispielwörter für deutsche Rechtschreibübungen. 7995 hochfrequente Wortformen der KAEDING-Zählung, rechnersortiert nach Phonem-Graphem-Beziehungen.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1976). *Hochfrequente deutsche Wortformen II. 7995 Wortformen der KAEDING-Zählung, rechnersortiert nach Wortartzugehörigkeit und Homographie.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1978). *Hochfrequente deutsche Wortformen IV. 7695/9566 Wortformen der KAEDING-Zählung, rechnersortiert nach Textsorten-Distribution.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1979). *Hochfrequente deutsche Wortformen III. 7995 Wortformen der KAEDING-Zählung, zu Grundformen zusammengefaßt und mit fünf neueren Häufigkeitslisten verglichen.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1980). *Sprechsilben im Deutschen: Typen, Häufigkeiten, Übungsbeispiele, rechnersortiert anhand von 7995 Wortformen der KAEDING-Zählung.* Herausgegeben vom Goethe-Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Ortmann, Wolf Dieter (1981). *Minimalpaare im Deutschen: Typen, Häufigkeiten, Übungsbeispiele, rechnersortiert anhand von 7995 Wortformen der KAEDING-Zählung; mit einem Anhang: Reimlexikon zur KAEDING-Wortliste.* Herausgegeben vom Goethe-

Institut, Arbeitsstelle für wissenschaftliche Didaktik, Projekt Phonothek. München: Goethe-Institut.

Schubenz, S. (1979). Eine Morphem-Analyse der deutschen Sprache und ihre lernpsychologische Bedeutung für die Vermittlung von Schriftsprachenkompetenz. In: Pilz, S., & Schubenz, S. [Hrsg.], *Schulversagen und Kindergruppentherapie. Pädagogisch-psychologische Therapie bei psychischer Entwicklungsbehinderung* (S. 239-255, 271-300). Köln: Pahl-Rugenstein.

Zipf, George Kingsley (1935/ 1968). *The Psychobiology of Language: An Introduction to Dynamic Philology.* Cambridge, Mass.: The M.I.T. Press.

Weitere Publikationen Kaedings

Dreinhofer, A., & Kaeding, F.W. (1886). *Unterrichtsbuch der Stolzeschen Stenographie.* Berlin: Mittler & Sohn. Teil II: *Übungs- und Lesebuch für Stolzesche Stenogramme.* Berlin: Mittler & Sohn 1886.

Kaeding, Friedrich Wilhelm (1884). *Erläuterungen zu dem Gedenkblatt Stolzescher Stenographen*. Berlin: Liebheit & Thiesen.

Kaeding, Friedrich Wilhelm (1889). *Die Denkmäler Stolze's*. Berlin: Selbstverlag. (= Stolze-Bibliothek I)

Kaeding, Friedrich Wilhelm (1889). *Die Denkmäler Stolze's*. Berlin: Selbstverlag. (= Stolze-Bibliothek I)

Kaeding, Friedrich Wilhelm (1889). *Über Stolze's Reden*. Berlin: Selbstverlag. (= Stolze-Bibliothek II)

Kaeding, Friedrich Wilhelm (1890/91). *Wilhelm Stolze's Briefwechsel*. Berlin: Selbstverlag. (= Stolze-Bibliothek III-VIII)

Kaeding, Friedrich Wilhelm (1891). *Biographie Wilhelm Stolze's*. Berlin: Selbstverlag. (= Stolze-Bibliothek IX/X)

Kaeding, Friedrich Wilhelm (1892). *Wilhelm Stolze's Arbeiten, seine Reden, Gutachten und wissenschaftlichen Vorträge. 3 Bde.* Berlin: Selbstverlag. (= Stolze-Bibliothek XI-XVIII)

Kaeding, Friedrich Wilhelm (1897). Vortrag über Geschichte, Wesen und Bedeutung der Stenographie. Steglitz bei Berlin.

Kaeding, Friedrich Wilhelm (1898). Über Geläufigkeitsuntersuchungen oder Feststellung der Schreibflüchtigkeit der Schriftzeichen. Steglitz bei Berlin: Selbstverlag.

Kaeding, Friedrich Wilhelm (1898). Der Vokal „e" im Einigungssystem Stolze-Schrey. *Magazin für Stenographie XIX, 487-490.*

Kaeding, Friedrich Wilhelm (1899). Zur Feststellung der Schriftzüge des Einigungssystems Stolze-Schrey. *Magazin für Stenographie XX, 75-78.*

Kaeding, Friedrich Wilhelm (1922). *Wilhelm Stolze, sein Leben und Wirken. 18 Bde in 1.* Magdeburg: Verlag für Stenographie.

Sekundärliteratur

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

Altmann, Gabriel (1985). Sprachtheorie und mathematische Modelle. *Christian-Albrechts-Universität Kiel, SAIS [= Seminar für Allgemeine und Indogermanische Sprachwissenschaft] Arbeitsberichte. H. 8, 1-13.*

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

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

Best, Karl-Heinz (2007). Helmut Meier (1897-1973). *Glottometrics 16, 122-124.*

Ito, Masamitsu (2005). Quantitative Linguistics in Japan. In: Köhler, Reinhard, Altmann, Gabriel & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (S. 82-95). Berlin/ New York: de Gruyter.

Wedegärtner, Elfriede [Bearb.] (1960). *Katalog der stenografischen Literatur (Bibliographie und Geschichte).* Hrsg. von der Bibliothek des Stenografischen Landesamtes Dresden. Dresden.

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

Biographische Informationen zu Kaeding

Böer, Oscar (1913). Unser Kaeding. *Der Deutsche Stenograph XIII, 295-298.*

Bonnet, Rudolf (1935). *Männer der Kurzschrift. 572 Lebensabrisse von Vorkämpfern und Führern der Kurzschriftbewegung.* Darmstadt: Winklers Verlag (Gebrüder Grimm).

Lambrich, Hans, & Kennerknecht, Aloys (1962). *Entwicklungsgeschichte der Deutschen Kurzschrift.* Darmstadt: Winklers Verlag – Gebrüder Grimm, S. 248, Portrait S. 23.

Nachruf [auf F.W. Kaeding]. *Der Deutsche Stenograph 28/ 1928, 129.*

Rechnungsrat F.W. Kaeding. *Der Deutsche Stenograph 8/ 1908, 412-413.*

Schneider, L., & Blauert, G. [Hrsg.] (1936). *Geschichte der deutschen Kurzschrift.* Wolfenbüttel: Heckners Verlag, S. 150.

Karl-Heinz Best, Göttingen

XXXVIII. Eduard Sievers (1850-1932)

Geb. 25.11.1850 in Lippoldsberg, gest. 30.3.1932 in Leipzig. Abitur in Kassel 1867; Studium der Klassischen Philologie, Germanistik und des Altenglischen in Leipzig und Berlin; Promotion 1870 in Leipzig. 1971 a.o. Prof. für germanische und romanische Philologie in Jena; 1876 Ordinarius; später Prof. für deutsche Philologie in Jena. 1883 Wechsel nach Tübingen; 1887 nach Halle als Prof. für deutsche Sprache und Literatur. 1892 Wechsel nach Leipzig. 1922 emeritiert. Sievers gehörte zu den Junggrammatikern und war ein Germanist im umfassenden Sinne: In seinen Publikationen werden Gegenstände vieler verschiedener germanischer Sprachen behandelt.

Sievers ist für die Quantitative Linguistik in zweierlei Hinsicht von Bedeutung: 1., weil er ein Prinzip benennt, aus dem später das Menzerath-Gesetz bzw. das Menzerath-Altmann-Gesetz erwuchs; 2., weil er bei vielen seiner Untersuchungen Statistiken erstellt. Im Zusammenhang mit Untersuchungen zur Lautquantität verweisen bereits Menzerath & de

Oleza (1928: 3ff.) mehrfach, aber nicht unkritisch auf Sievers. Die Entwicklung des Prinzips lässt sich aus den verschiedenen Auflagen des entsprechenden Buches von Sievers ablesen:
In der ersten Auflage *Grundzüge der Lautphysiologie* (Sievers 1876: 122) findet sich dazu folgendes: „U e b e r l a n g e V o c a l e (resp. Diphthonge) stehen unter dem Einflusse des Accentes statt gewöhnlicher Längen häufig in einsilbigen Worten in Pausa, d.h. solchen, denen nicht mehr eine zu demselben Satze gehörige Silbe folgt. Mehrsilbige Formen desselben Wortes zeigen dann einfache Länge; man vgl. also etwa *to̲t* und *to̲tə*[8],..., namentlich in Mundarten, welche den geschliffenen Accent... besitzen.“ Also auch schon in der ersten Auflage wird die Länge der Silben eines Wortes bzw. ihrer Vokale in Bezug gesetzt zur Zahl der Silben dieses Wortes.

Die zweite Auflage, von da an mit dem Titel *Grundzüge der Phonetik*, ist etwas ausführlicher: „L a n g e Silben werden zu überlangen... Für die Praxis ist hier wieder auf die schon...berührte Neigung mancher Sprachen hinzuweisen, lange Monosyllaba in Pausa (d.h. am Satzende) oder bei starkem Nachdruck zu überlangen Silben zu machen. In dem einsilbigen *t̲o̲t̲* ist nicht nur der Vocal länger als in dem zweisilbigen *to̲tə*[9], sondern auch die Pause zwischen Verschluss und Oeffnung des *t* wird gedehnt...“ (Sievers 1881: 194f.).
Auf die einschlägige Formulierung dieses Prinzips in der fünften Auflage verweisen Altmann & Schwibbe (1989: 60). Das entsprechende Zitat kommt zuerst in der vierten Auflage von Sievers *Grundzüge der Phonetik* vor: „Vor Allem aber regelt sich die Silbendauer zu einem grossen Theile nach der S i l b e n z a h l d e r S p r e c h t a k t e, denen die betreffenden Silben angehören... Sprechtakte, die an äusserem Umfang, d.h. eben an Silbenzahl, nicht zu verschieden sind, werden gern mit gleicher oder doch annähernd gleicher Dauer gesprochen..., vgl. etwa Sprechtakte wie *heil*, | *heilig*, | *heilige*, | *heiligere* | u. s. w. Dann entfällt aber auf jede Einzelsilbe eines aus weniger Silben bestehenden Sprechtakts ein grösseres Stück Zeit als auf die Einzelsilben eines Taktes von mehr Silben. Aber auch selbst da, wo Gleichheit der Dauer der Sprechtakte nicht erreicht wird, herrscht doch stets die Neigung, vielsilbige Takte schneller, solche mit weniger Silben langsamer zu sprechen, d.h. eben die Silbendauer nach der Taktform zu modificiren“ (Sievers 1893: 240f.). Für die vierte Auflage 1893 kann man konstatieren, dass Sievers das Prinzip „Je mehr Silben ein Sprechtakt hat, desto kürzer werden die Silben gesprochen“ erkannt hat. Vor dieser vierten Auflage wird dieser Zusammenhang nicht ganz so klar und nicht ganz so allgemein vorgetragen.

Sievers hat für dieses Prinzip jedoch mindestens einen Vorläufer: den Romanisten Diez, auf den Grégoire (1899: 161) hinweist: „Wie in den Schwestersprachen kürzt sich die Länge des Stamm- oder Tonvocals, wenn durch Ableitung oder Flexion der Ton fortrückt...“ (Diez 1856: 467). Zu weiteren Stationen der Entwicklung des Menzerathschen Gesetzes: Menzerath & de Oleza (1928: 3-8) sowie Altmann & Schwibbe (1989: 37ff., zu Laut- und Silbendauer bes. 60-64).

Weiterhin sei darauf hingewiesen, dass Sievers in seinen Vers-Untersuchungen vielfach Statistiken erhebt, um die Häufigkeit verschiedener Typen zu dokumentieren (z.B. in Sievers 1885/ 87). In Sievers (1879: 356) gibt er eine Statistik zur Länge von Grundversen in dem *ljóðaháttr* (eine besondere Art der Strophenbildung, Vries 1964: 25) wieder, die zwischen 2 und 6 Silben lang sein können. Man kann dies als einen Fall von Diversifikation betrachten (Altmann 1991, 2005) und nach einem dafür geeigneten Modell suchen. Es hat sich erwiesen, dass man an Sievers‘ Beobachtungsdaten die verschobene Conway-Maxwell-Poisson-Verteilung

[8] Im Original steht ein <e> mit untergesetztem <°>.
[9] S. Fußnote 1.

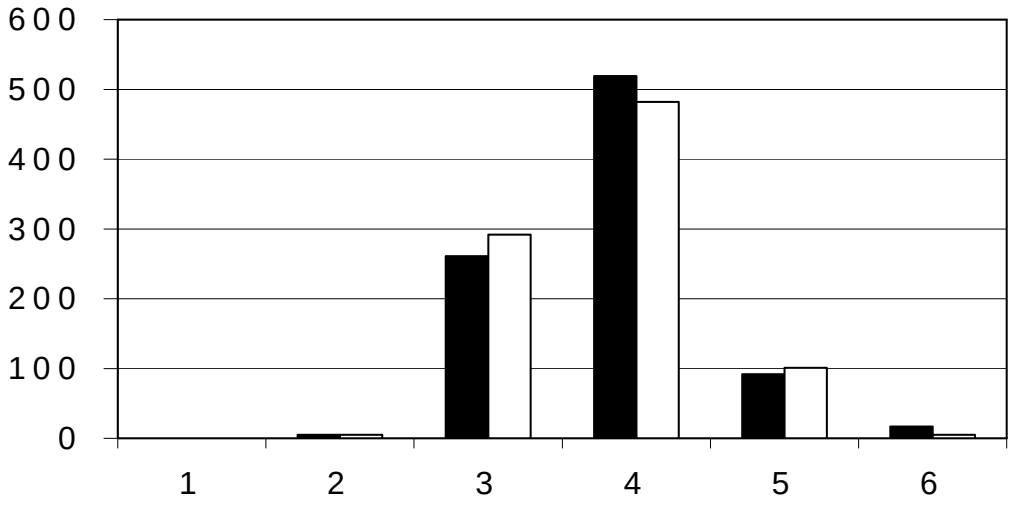
$$P_x = \frac{a^{x-2}}{(x-2)!^b} C, \qquad x = 2,3,4,\ldots$$

anpassen kann. Das Ergebnis ist in Tabelle (1) wiedergegeben:

Tabelle 1
Anpassung der Conway-Maxwell-Poisson-Verteilung

x	n_x	NP_x
2	5	5.21
3	261	291.87
4	510	481.74
5	92	101.18|
6	17	5.00|
a = 1.2355 b = 0.6719 C = 0.0057		

Legende zur Tabelle: x: Zahl der Silben pro Vers; a und b sind die Parameter der Conway-Maxwell-Poisson-Verteilung; C ist der Diskrepanzkoeffizient, der mit $C \leq 0.01$ eine gute Anpassung der Verteilung an die beobachteten Daten anzeigt. Die senkrechten Striche in der Tabelle weisen darauf hin, dass diese beiden Längenklassen zusammengefasst wurden.



Graphik zu Tabelle 1

Man sieht, dass die Übereinstimmung recht gut ist, wie ja auch schon der Diskrepanzkoeffizient bestätigt hat.

Sievers gehört mit seinen Ausführungen zur Abhängigkeit der Lautlänge von der Wortlänge zu den Vorläufern des Menzerath-Altmann-Gesetzes und damit zu den Autoren, die der Quantitativen Linguistik bei ihrer Suche nach Sprachgesetzen den Weg bereitet haben. Darüber hinaus sind seine statistischen Erhebungen zur Poesie offenbar erst noch zu entdecken.

Auf ausführlichere Darstellungen von Leben und Werk e. Sievers wird verzichtet; seine Werke sind von Karg-Gasterstädt (1933) vollständig erfasst; weitere Informationen zu seinem Leben sind Frings (1933) und Erhard (1996) sowie der unten angegebenen Internetadresse zu entnehmen.

Literatur

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

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

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

Diez, Friedrich Christian (1856-60). *Grammatik der romanischen Sprachen. Erster Theil.* 2. neu verfasste Auflage. Bonn: Eduard Weber.

Grégoire, Antoine (1899). Variations de durée dans la syllable française. *La parole 1, H. 3, 4 und 6.*

Menzerath, Paul, & de Oleza, Joseph M. (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung.* Berlin/ Leipzig: de Gruyter.

Sievers, Eduard (1876). *Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1879). Beiträge zur Skaldenmetrik II. *Beiträge zur Geschichte der deutschen Sprache und Literatur V/, 265-376.*

Sievers, Eduard (1881). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Zweite wesentlich umgearbeitete und vermehrte Auflage der „Grundzüge der Lautphysiologie“. Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1885). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* 3., verbesserte Auflage. Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1885/ 87). Zur Rhythmik des germanischen Alliterationsverses. *Beiträge zur Geschichte der deutschen Sprache und Literatur X/ 1885, 209-220, 220-314, 451-545; XII/ 1887, 454-482, 482-491, 492-497, 498-503.*

Sievers, Eduard (1893). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* 4., verbesserte Auflage. Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1901). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* 5. verbesserte Auflage. Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1912). Zur älteren Judith. In: Sievers, Eduard, *Rhythmisch-melodische Studien. Vorträge und Aufsätze* (S. 112-141). Heidelberg: Winter. (Zuerst 1908 in Prager deutsche Studien 8, 179ff.)

Vries, Jan de (1964). *Altnordische Literaturgeschichte. Bd. 1.* Berlin: de Gruyter.

Quellen zu Biographie und Bibliographie

Ehrhardt, Horst (1996). Eduard Sievers. In: Stammerjohann, Harro (ed.), *Lexicon gramma-*

ticorum. Who's Who in the History of World Linguistics (S. 860-861). Tübingen: Niemeyer.

Frings, Theodor (1933). Eduard Sievers, geboren am 25. November 1850, gestorben zu Leipzig am 30. März 1932. In: *Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Philologisch-historische Klasse, 85. Bd., 1. Heft* (S. 1-56). Leipzig: Hirzel. Auch in: Sebeok, Thomas A. (ed.) (1966). *Portraits of Linguists. A Biographical Source Book for the History of Western Linguistics, 1746-1963. Vol. II (S. 1-52).* Bloomington/ London: Indiana University Press.

Karg-Gasterstädt, Elisabeth (1933). Schriftenverzeichnis. In: *Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Philologisch-historische Klasse, 85. Bd., 1. Heft* (S. 57-92). Leipzig: Hirzel. Auch in: Sebeok, Thomas A. (ed.) (1966). *Portraits of Linguists. A Biographical Source Book for the History of Western Linguistics, 1746-1963. Vol. II. From Eduard Sievers to Benjamin Lee Whorf* (S. 1-52). Bloomington/ London: Indiana University Press.

http://www.catalogus-professorum-hallensis.de/sieverseduard.html

Karl-Heinz Best, Göttingen

XXXIX. Ferdinand Schrey (1850-1938)

Geb. 19.7.1850 in Wuppertal-Elberfeld, gest. 2.10.1938 in Berlin. Nach einer Banklehre und der Teilnahme am deutsch-fränzösischen Krieg 1870/71 arbeitete er als kaufmännischer Angestellter. Er wurde zuerst Teilhaber, später Alleininhaber einer Knopffabrik in Wuppertal-Barmen. Schrey befasste sich als Anhänger von Gabelsberger mit der Kurzschrift und entwickelte ein eigenes, einfacheres System. 1885 begann er mit dem Vertrieb von Schreibmaschinen. 1891 gründete er ein eigenes Schreibmaschinengeschäft in Berlin; es folgten ein Kurzschriftverlag und eine Einrichtung für den Unterricht von Kurzschrift und Maschinenschreiben. Schrey ist einer der Namensgeber für das Kurzschriftsystem Stolze-Schrey. Der Ausdruck „Stenotypistin“ wird ihm zugeschrieben.

Auf Schrey macht Kaeding (1897/98: 39) aufmerksam. Für die Quantitative Linguistik ist er deshalb von Interesse, weil er noch vor Kaeding (1897/98) eine Statistik der Laute veröffentlicht, von der er mitteilt, dass sie „vor längeren Jahren von Herrn Lehrer Heinrich Heine in Essen an der Ruhr auf Grund von 50000 Silben zusammenhängenden Stoffes aufgestellt“ (Schrey 1891: 6) worden sei. Ausgewertet wurden „Reden aus Abgeordnetenhaus und Reichstag“ (ebenda). Die Statistik ist recht differenziert nach den Positionen der Laute im Wort; sie wird ergänzt durch statistische Angaben zur Häufigkeit von Vorsilben und Endungen sowie zu Lautkombinationen in An- und Auslaut (Schrey 1891: 6f.).

Hier wird exemplarisch nur die Rangordnung der Einzellaute, unabhängig von ihrer Position im Wort, betrachtet. Bei entsprechenden Tests hat sich ergeben, dass an diese Daten Altmanns Modell (1993: 62) für beliebige Rangordnungen

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

das keine Verteilung, sondern eine Folge darstellt, mit sehr guten Ergebnissen angepasst werden kann. (Das Modell hat sich schon mehrmals für ähnliche Rangordnungen bewährt

(Best 2004/5; 2005; 2006: 57-60), u.a. bei der 100000-Laute-Zählung“, die Meier (1967: 249ff.)

Tabelle 1
Anpassung des Modells für beliebige Rangordnungen
an die von F. Schrey mitgeteilten Lauthäufigkeiten (Vokale)

Rang	Lautbe-zeichnung	n_x	NP_x	Rang	Lautbe-zeichnung	n_x	NP_x
1	e	19999	19999.00	8	ü	1020	1134.35
2	i	9405	10390.84	9	ä	723	893.11
3	a	7002	6159.79	10	ö	426	717.19
4	u	4141	3983.38	11	eu	355	585.62
5	o	3000	2740.38	12	ai	54	485.09
6	ei	2791	1974.70	13	äu	38	406.85
7	au	1036	1475.21	14	y	10	344.96
a = 3.5595			b = 0.8885		D = 0.9918		

Legende zu den Tabellen:

n_x: beobachtete Häufigkeit der betreffenden Einheit;

NP_x: aufgrund des Modells berechnete Häufigkeit der betreffenden Einheit;

a, b, c: Parameter des Modells von Altmann; bei den Berechnungen wird $c = y_1$ gesetzt.

D: Determinationskoeffizient. Der Determinationskoeffizient ist akzeptabel, wenn $D \geq 0.80$, und sehr gut mit $D \geq 0.90$. In diesem Fall wurde also mit $D = 0.99$ ein exzellentes Testergebnis erzielt.

Aus der Tabelle wird deutlich, dass es sich nicht um eine reine Lautstatistik handeln kann, sondern auch um die Schreibweise der Laute. Anders wäre das Nebeneinander von <ei> und <ai> sowie von <eu> und äu> nicht zu erklären. Auch die fehlende Berücksichtigung der Vokallänge fällt ins Auge. Zielsetzung des Stenographen ist denn auch die „beste[...] Verteilung der Zeichen auf die Laute“ (Schrey 1891: 6). Den Hintergrund dafür bilden die Auseinandersetzungen zwischen den Stenographenschulen um das bestmögliche Kurzschriftsystem.

Tabelle 2
Anpassung des Modells für beliebige Rangordnungen
an die von F. Schrey mitgeteilten Lauthäufigkeiten (Konsonanten)

Rang	Lautbe-zeichnung	n_x	NP_x	Rang	Lautbe-zeichnung	n_x	NP_x
1	n	14960	14960.00	14	z	1939	1525.05
2	r	11227	11766.05	15	k	1610	1344.07
3	s	8120	9399.48	16	v	1398	1189.84
4	t	8072	7612.40	17	ng	1040	1057.68
5	d	7341	6240.21	18	sch	988	943.83
6	l	4962	5170.86	19	ss[10]	818	845.30
7	g	4081	4326.43	20	p	757	759.62
8	ch	3450	3651.62	21	j	302	684.82
9	m	3118	3106.55	22	ck	151	619.23

[10] Hier steht in der Vorlage die Kombination aus langem, geschweiftem <s> und einfachem <s>.

10	w	2742	2661.94	23	x	24	561.52
11	b	2731	2296.05	24	y	10	510.56
12	f	2231	1992.46	25	c	5	465.40
13	h	2016	1738.67	26	qu	4	425.26
a = 14.2679			b = 10.7947		D = 0.9850		

Auch hier wird wieder deutlich, dass es sich nicht um eine reine Lautstatistik handelt: So wäre zu fragen, welcher Lautunterschied zwischen <k> und <ck>, welcher zwischen <s> und <ss> besteht, wenn es nur um Laute gehen sol; außerdem, wo der Unterschied zwischen stimmhaftem und stimmlosem <s> bleibt.

Tabelle 3
Anpassung des Modells für beliebige Rangordnungen
an die von F. Schrey mitgeteilten Lauthäufigkeiten (alle Laute)

Rang	Lautbe-zeichnung	n_x	NP_x	Rang	Lautbe-zeichnung	n_x	NP_x
1	e	19999	19999.00	21	k	1610	1519.02
2	n	14960	15686.65	22	v	1398	1413.33
3	r	11227	12625.46	23	ng	1040	1318.21
4	i	9405	10375.37	24	au	1036	1232.31
5	s	8120	8673.89	25	ü	1020	1154.48
6	t	8072	7356.59	26	sch	988	1083.74
7	d	7341	6316.25	27	ss	818	1019.25
8	a	7002	5480.55	28	p	757	960.31
9	l	4962	4799.30	29	ä	723	906.30
10	u	4141	4236.76	30	ö	426	856.69
11	g	4081	3766.97	31	eu	355	811.00
12	ch	3450	3370.66	32	j	302	768.85
13	m	3118	3033.31	33	ck	151	729.88
14	o	3000	2743.84	34	ai	54	693.77
15	ei	2791	2493.61	35	äu	38	660.25
16	w	2742	2275.86	36	x	24	629.09
17	b	2731	2085.23	37	y[11]	10	600.06
18	f	2231	1917.42	38	y	10	572.99
19	h	2016	1768.92	39	c	5	547.69
20	z	1939	1636.90	40	qu	4	524.01
a = 7.5278			b = 5.4734		D = 0.9836		

In allen drei Fällen erhält man also ein hervorragendes Ergebnis für den Determinationskoeffizienten, so dass die Anpassung des Altmannschen Modells für beliebige Rangordnungen als erfolgreich betrachtet werden kann.

Schrey gehört in die Annalen der Quantitativen Linguistik, weil er noch vor Kaeding (1897/98) eine statistische Erhebung zu Laut-Buchstaben-Häufigkeiten zugänglich gemacht hat. Wann diese genau entstanden ist, wird allerdings nicht mitgeteilt. Mit der Auswertung

[11] <y> steht sowohl in der Tabelle der Vokale als auch in der für die Konsonanten mit gleicher Häufigkeitsangabe. Es ist nicht klar, ob es sich um die gleichen oder verschiedene Vorkommen von <y> handelt.

von 50000 Silben bietet er verglichen mit Nowak (1848), der lediglich 1000 Laute ausgezählt hatte, eine wesentlich bessere Datenbasis.

Literatur

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

Best, Karl-Heinz (2004/5). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/ 11, 21-32.*

Best, Karl-Heinz (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 231, 119-127.*

Best, Karl-Heinz (2006). Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71.*

Meier, Helmut (1967). *Deutsche Sprachstatistik.* Zweite erweiterte und verbesserte Aufl. Hildesheim: Olms.

Nowak, Josef (1848). *Leicht lesbare Geschwindschrift (Tachygraphie, Stenographie), oder: Ausführliche Anleitung zum Selbstunterrichte in der Kunst, so schnell zu schreiben, als ein öffentlicher Redner spricht. Für alle Stände.* Dritte, umgearbeitete Auflage. Wien: Sallmayer und Comp.

Schrey, Ferdinand (1891). *Das stenographische Zeichenmaterial und seine Verwendung.* Vortrag auf dem IV. Internationalen Stenographentag zu Berlin, Anfang Oktober 1891. Berlin: Selbstverlag von F. Schrey.

Quellen zu Schrey

Bonnet, Rudolf (1935). *Männer der Kurzschrift. 572 Lebensabrisse von Vorkämpfern und Führern der Kurzschriftbewegung.* Darmstadt: Winklers Verlag (Gebrüder Grimm).

Kaeding, Friedrich Wilhelm [Hrsg.] (1897/98). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Beiheft zu Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.*

Schneider, L., & Blauert, G. [Hrsg.] (1936). *Geschichte der deutschen Kurzschrift.* Wolfenbüttel: Heckners Verlag, S. 180ff.

Webseite: http://www.steno.ch/htm/120.htm#biografien (Webseite des Schweizerischen Stenographenverbandes SSV).

Karl-Heinz Best, Göttingen

XL. Heinrich August Kerndörffer (1769-1846)

Geb. 16.12.1769 (Leipzig), gest. 23.9.1846 (Reudnitz bei Leipzig). Studium der Philosophie in Leipzig; dort Lehrer an der Nicolai-Schule und Universitätsdozent für deutsche Sprache und Deklamation und Privatgelehrter; seit 1805 Mitglied und zeitweise „Meister vom Stuhl" der Leipziger Freimaurerloge „Apollo". Autor von Kinderbüchern und vielen Trivialromanen verschiedener Genres (Weidemeier 1967), eines Zauberbuchs, aber auch von Schriften zu Deklamation, Freimaurerei und Pädagogik. „Ein Erfolgsschriftsteller zu Leb-

zeiten, kennen ihn spätere Literaturgeschichten nur noch als obskuren Trivialautor“ (Meyer-Kalkus 2001: 62). Kerndörffer ist in der Literaturwissenschaft dafür bekannt, dass er Deklamationslehrer von Heinrich von Kleist war (Kleist in einem Brief vom 13.3.1803 an Ulrike von Kleist). Für seine Arbeit als Deklamationslehrer verdiene Kerndörffer „immerhin eine Teilrehabilitierung“, meint Meyer-Kalkus (2001: 62).

Kerndörffer ist für die Quantitative Linguistik erwähnenswert, da er womöglich der erste Autor ist, der Angaben zur Häufigkeit von Buchstaben im Deutschen liefert. Seinen Angaben müssen Zählungen zugrunde liegen; allerdings präsentiert er keine Statistik, sondern lediglich Häufigkeitsklassen ohne genaue Zahlenangaben. Auf Kerndörffer wird man von Kaeding (1897/98: 38) aufmerksam gemacht: „Angaben über Häufigkeit der Buchstaben befinden sich in dem Werke über K r y p t o g r a p h i e, von K e r n d ö r f f e r: Leicht faßliche Anleitung zur Kryptographie 1835. (In der deutschen Sprache: e, i, a, o, u; im Lateinischen und Spanischen o am häufigsten. Konsonanten im Deutschen: n, t, r, s, c, d, h, m, v, w, b, g, f, k, z, p, q, x).“ [12] Die entsprechenden Angaben finden sich in Kerndörffer (1835: 98 für Vokale und 101 für Konsonanten). Er ordnet die Konsonanten nach Häufigkeitsklassen und setzt sie in Beziehung zu den Vokalen. Er gruppiert als häufigste Klasse: n, t, r, s; danach c, d, h, l (das Kaeding übersehen hat); es folgen die übrigen Klassen wie von Kaeding aufgelistet, der allerdings den nach Kerndörffer seltensten Buchstaben y auslässt. Ein Vergleich mit anderen Zählungen (z.B. Best 2005a, b) zeigt, dass Kerndörffer die Häufigkeitsrangordnung im Prinzip erkannt hat. Außerdem teilt Kerndörffer Beobachtungen zur Position und Distribution der Buchstaben mit. Bleibt zu ergänzen, dass in Kerndörffer (1835: 99) auch noch Hinweise zu Französisch, Italienisch, Lateinisch und Spanisch zu finden sind.
Für die Historie der Quantitativen Linguistik ist also festzustellen, dass Buchstabenzählungen schon für 1835 angesetzt werden können, auch wenn deren Ergebnis und genaue Datenbasis nicht mitgeteilt werden. Die erste veröffentlichte Statistik stammt vermutlich von Förstemann (1846; vgl. dazu Best 2006); die nächste bisher bekannte Zählung findet sich bei Nowak (1848), der eine Statistik, die auf der Auszählung von 1000 Buchstaben beruhte, ermittelte und mitteilte.

Literatur

Best, Karl-Heinz (2005a). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 231, 119-127.*

Best, Karl-Heinz (2005b). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11, 9-31.*

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

Förstemann, Ernst (1846). Ueber die numerischen Lautverhältnisse im Deutschen. *Germania 7, 83-90.*

Kerndörffer, Heinrich August (1835). *Leicht faßliche Anleitung zur Kryptographie oder den verschiedenen Arten der geheimen Schreibekunst, in Verbindung mit der Stenographie und Tachygraphie oder Geschwindschreibekunst und ihrer Anwendung für die mannichfaltigen Verhältnisse und Angelegenheiten des Staatslebens neuerer Zeit.* Leipzig: Verlag von L. Fort.

Kleist, Heinrich von (1999). Brief vom 13. März 1803 an Ulrike von Kleist. In: *Heinrich von Kleist, Sämtliche Werke* (Brandenburger Ausgabe), hrsg. von Roland Reuß und Peter Staengle. 4: Briefe, Bd. 2 (S. 243-251). Basel/ Frankfurt: Stroemfeld.

[12] Das Buch ist in Bibliotheken nicht oft vertreten; man findet es im Bestand der Technischen Informationsbibliothek der Universität Hannover, Haus 2, Rethen.

Meyer-Kalkus (2001). Heinrich von Kleist und Heinrich August Kerndörffer. Zur Poetik von Vorlesen und Deklamation. In: *Kleist-Jahrbuch 2001* (S. 55-88). Stuttgart/ Weimar: Metzler.

Nowak, Josef (1848). *Leicht lesbare Geschwindschrift (Tachygraphie, Stenographie), oder: Ausführliche Anleitung zum Selbstunterrichte in der Kunst, so schnell zu schreiben, als ein öffentlicher Redner spricht. Für alle Stände.* Dritte, umgearbeitete Auflage. Wien: Sallmayer und Comp.

Weidemeier, Hartmut (1967). *Heinrich August Kerndörffer. Untersuchungen zum Trivialroman der Goethezeit.* Diss., Bonn.

Quellen zu Kerndörffer

Brümmer, Franz (1876). *Deutsches Dichter-Lexikon: biographische und bibliographische Mittheilungen über deutsche Dichter aller Zeiten; unter besonderer Berücksichtigung der Gegenwart. Band 1.* Eichstätt (u.a.): Krüll.

Kaeding, Friedrich Wilhelm [Hrsg.] (1897/98). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Beiheft zu Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.*

Killy, Walther (Hrsg.) (1990). *Literatur Lexikon. Autoren und Werke deutscher Sprache. Bd. 6.* Gütersloh/ München: Bertelsmann Lexikon Verlag.

Rupp, Heinz, & Lang, Ludwig [Hrsg.] (1981). *Deutsches Literatur-Lexikon. Biographisch-bibliographisches Handbuch*, begründet von Wilhelm Kosch. Dritte, völlig neu bearbeitete Auflage. Achter Band. Bern/ München: Francke.

Karl-Heinz Best, Göttingen

Book Reviews

Jeehyeon Eom, *Rhytmus im Akzent. Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhytmus im Russischen*. München: Verlag Otto Sagner 2006. 200 pp.
Reviewed by **Ján Mačutek.**

The book, consisting of four chapters, is dedicated to investigations of rhythm in Russian accent. Rhythm, understood as a regular alternation of accentuated and non-accentuated syllables, is modeled by several mathematical tools. All models are tested on 26 Russian texts (3 poems, 13 short prose texts, 10 samples from longer prose texts). Accentuated and non-accentuated syllables are represented by a binary code (i.e., by the numbers 1 and 0 respectively).

In the first two chapters the author presents a very thorough discussion on basic notions connected with the investigated topic. Although research in this area has more than 300-year-long history (some pioneer works are mentioned in the book), until recently it lacked clear definitions and many aspects were left to linguists' intuition. Needless to say, in such a situation the discussion which brightens the old fuzziness and leads to more precise understanding of accent phenomena must be appreciated. The author strictly differentiates between the notions of accent ('*Akzent*', an abstract entity of language) and stress ('*Betonung*', a realization of accent). We note that there is no difference between them in Russian terminology (which uses the word '*уда́рение*' for both of them). As is obvious from the title, the study concentrates on accent.

Then, in the third chapter (the fourth one consists of final remarks), the author proceeds to mathematical models and their evaluation. The aim is to model accent regularities, which can be done in several ways. The one which is used as the first is a model for length of accent groups. Thanks to the binary code, quantification is straightforward, we take the number of non-accentuated syllables between two accentuated syllables (with several possible minor modifications, e.g., to include or not to include the accentuated syllables, etc.).

Two models are suggested, namely the chaos model (‘*Chaosmodel*’) and the brake mechanism model (‘*Bremsmechanismusmodell*’). The first of them is an application of a model derived by Strauss et al. (1984). In our context it is based on the assumption that throughout the text (i.e., for every position) the probabilities of occurrence of accentuated and non-accentuated syllable are the same. The model does not yield a good fit for 10 out of 26 texts. For poems, its failure was expected. For prose texts, it is stated that the general hypothesis of chaotic distribution of accent was wrong. It can be true, but we point out another weakness of the model which can perhaps also be a reason why the model does not fit data in some cases. While some text entities can occur in constant sequences (like 1111 or 000), within the author's approach a word can bear maximally one accent (i.e., nothing like secondary accent is considered). The assumption of the same probability of accentuated and non-accentuated syllable makes sense for short (one- or two-syllable) words. Once we have, e.g., a four-syllable word, the position of the accentuated syllable can be considered random, but in the word there must be three non-accentuated syllables (but the model allows all of them to be accentuated). Of course, no model can take into account everything and neglecting is allowed, but if a text contains 'too many too long' words, the discrepancy between model assumptions and reality can play its role. It would be interesting to introduce a parameter representing word length (mean? maximum?) into the model (and not only into this one).

The brake mechanism model is linguistically justified (a tendency to split 'too long' sequences of non-accentuated syllables) and the author provides its honest mathematical

derivation, leading to the extended binomial distribution. The model is not general, as it is not adequate for poems (again no surprise, the rhythmic organization of a poetic text has a higher priority than its 'natural look'), but it fits data almost perfectly for prose texts (there is only one exception among 23 analyzed texts). However, there is a question to be answered. The parameter *n* of the extended binomial distribution is interpreted as the highest number of non-accentuated syllables between two neighboring accentuated syllables in the given text. The values it attains are unrealistic in some cases (*n* = 77 in Text 3 and Text 15, *n* = 88 in Text 23).

The next model, the autocorrelation, is used to model linear dependencies between individual terms of a sequence. A relatively high autocorrelation between relatively distant terms means (in our context) that rhythmic motives are repeated (without any further specification of the motives, i.e., it only tells us whether rhythmic motives are repeated or not). Here we do not test the adequacy of the model, autocorrelation can be applied to any sequence. One can again easily see the difference between the rhythm of poetic and prosaic texts – in poetic texts at least the first six correlations are significant, whereas none of the prosaic texts has more than three significant correlations.

The staircase model (‘*Treppenmodell*’) is applied to cumulative frequencies of accentuated syllables (*x* is the position of a syllable in a text, *f*(*x*) the number of accentuated syllables from the beginning to the *x*-th position). From among a huge number of possible functions the author chooses the most simple one, namely $f(x) = ax$ (i.e., the linear function with the intercept zero). As the number of accentuated syllables is, naturally, always an integer, the function is modified to the shape $f(x) = \mathrm{int}(ax)$, int(*x*) being the floor function. The goodness of fit, measured by the determination coefficient, is very convincing. The parameter *a* is described as ‘*Akzentuierungspotenzial*’, i.e., potential of being accentuated. We recall once more word length here (cf. Best 2005 and Grzybek 2006 for studies on word length). If we define word phonologically and merge clitics with the words to which they phonologically ‘belong’, we obtain $a = 1/\overline{L}$, $\overline{L}$ being the mean word length (which means that cumulative frequencies of accentuated syllables in Russian, although the Russian accent has no fixed position in the word, depend only on word length, hence the model should be applicable also to other languages). Of course, this hypothesis must be tested on many texts before it can be considered corroborated.

Markov chains (of order 1, 2 and 3) are applied to model probabilities of transition from one state to another (as far as accent is considered, there are only two possible states, namely accentuated and non-accentuated syllables). Then the distance between two states of the same type (between two neighboring 1’s in our case) is modeled by the apparatus. The Markov chain of order 3 yields a satisfactory fit for all texts but one. Even order 2 is enough for 10 texts. Problems are connected mainly with parameter interpretation – the higher the order of the Markov chain, the greater the number of parameters in the model. Also the order itself remains unexplained for the time being. We only remind that it coincides with the mean word length, which – we guess – will be somewhere between 2 and 3 in most of the analyzed texts.

Finally, the stationary distribution (the limit of the transition matrix) is found for all texts and for all three Markov chains orders. Comparisons of the numerically obtained limits, e.g., with the ‘ideal dactylic sequence’ (100100100…) could perhaps be used as a method for characterization and/or classification purposes.

Eom’s book is an important step forward in the research of accent rhythm. He suggests several models and does his best in their mathematical derivation as well as linguistic interpretation, he fits them to data and he does not hide weak points or gaps in his study. A lot of work remains to be done, but this book traces a promising way.

References

Best, K.-H. (2005). Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 260-273.* Berlin/New York: de Gruyter.

Grzybek, P. (ed.) (2006). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht: Springer.

Strauss, U., Sapok, Ch., Diller, H.J., Altmann, G. (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7, 73-100.*

Haruko Sanada, *Investigations in Japanese Historical Lexicology* (Revised Edition). (= Göttinger Linguistische Abhandlungen 6). Göttingen: Peust & Gutschmidt Verlag 2008, 210 pp. ISBN 978-3-933043-12-2. Reviewed by **L. Uhlířová.**

If anybody tempted to judge on the content of Sanada´s monograph exclusively by its title, he might consider the topic to be so special and narrow that it could be properly understood only by Japanologists. However, it is not so. Sanada does not investigate etymology. She is not interested in the fate of individual words; her book cannot be listed under the heading of "descriptive historical lexicography". She does not even expect her readers to speak Japanese or have any knowledge about the three different kinds of script used for Japanese, hiragana, katakana and kanji. Her aims are much broader. She is interested in **general mechanisms** which control the dynamicity of the word stock development, i.e. the processes of the birth, growth or decay, spread or disappearance of words in the word stock in a concrete language during a historical period. For Sanada, Japanese is a **case**, a concrete **instance** to demonstrate how the general mechanisms work. The main idea and aim, which runs through the whole book like a golden thread, is to show that "idiosyncrasies can be revealed only if a background regularity is known" (p. 166)

And she has succeeded. She took inspiration from the synergetic linguistics, which was proposed, elaborated and applied in the famous books and studies by G. Altmann, K.-H. Best, R. Köhler, G. Wimmer and others (there is a long list of references in the end of the book), and she has come to pioneering results.

Naturally, it would be hardly possible to go through the dynamics of the **whole** word stock in a single book by a single author. That´s why Sanada chose a limited set of words, called **scholarly terms,** to start with. Why scholarly terms? In the second half of the 19th century (during the Meiji era, 1868–1912) Japan was opened up to massive cultural exchange with Western countries and an invasion of the European culture also involved language. Many new words = scholarly terms for abstract political and cultural concepts, as well as words for objects used in everyday life and other "civilization elements" were introduced in the form of kango (= Chinese-character words). A long-lasting period of very intensive, dynamic development of the word stock started. Some terms took root and spread, some other disappeared later. Sanada investigates the dynamicity of the processes from the very beginning in the 19th century till the end of the 20th century. The time span exceeding a whole century is, undoubtedly, a great task. The main material and the starting point was the Tetsugaku Jii (= Dictionary of Philosophy), a crucial lexicographic work of the late 19th century, published in three different editions; then fourteen different bilingual dictionaries (mainly: Japanese–English ones) were compared; and finally, an extensive and detailed comparison of the word stock of the present-day media (including television programs,

newspapers and magazines) and novels of the 20th century followed.

Although in the Introduction Sanada promises to use "elementary mathematics", it is exactly the mathematics that allows her to open relevant questions and to find satisfactory answers. Firstly, with the help of absolute and relative frequencies she showed the general tendencies of the headword changes in the three editions of Tetsugaku Jii; she attested which words survived, which were added and which were deleted, taking into account the fact that the third edition was over five times larger and thirty years "younger" than the first one. Then, an application of Wimmer-Altmann´s unified model of "rate of change" helped her to reveal similarities and differences between fourteen dictionaries (and differences between two "schools of dictionaries") as well as to describe the speed of the spread of the terms. Statistical correlations between frequency lists (tested by chi-square tests) proved not only individual, generational and other differences in the vocabulary of post-war media, but also a significant fact (very probably, of a universal linguistic nature) concerning the highest frequency classes: Some words introduced once as very abstract specialist terms now form a part "of the central core of vocabulary essential for everyday life as words expressing present-day abstract concepts" (p. 88). The regression analysis showed that scholarly terms share quantitative properties that match types of functions (and types of distributions) which have been attested already for other languages. E.g., there is a correlation between word length (measured in number of morae, syllables and strokes) and number of meanings, between word length and word frequency, as well as between word frequency and number of meanings, etc.

To sum up:

Firstly, Sanada is the first who has brought convincing evidence that the general mechanisms of the dynamic changes in the word stock on time axis hold good for Japanese. She has written the first book of this kind in the Japanese linguistics.

Secondly, having presented original data and having tested them successfully, she gives a challenge to other lexicologists to follow her and to compare her results with data from other languages – and thus, hopefully, to bring more proofs of the explanatory power of the synergetic theory of language laws. Of course, Sanada is well aware of the fact that such language laws (or, candidates for them) are "statements about mechanisms, whose results are usually fuzzy phenomena not similar to those in macrophysics" (p.2) and that they very probably will vary from one language to another. The more inspiration may be found in her monograph. After all, similar (or, at least comparable) highly dynamic periods in the vocabulary development may be easily evidenced in many other (if not in all) languages of the world. Let us consider, e.g., the strong influence of Latin and Greek on European languages during the Renaissance, or the mutual influence of Slavic languages during the National Revival, or the massive impact of English(es) all over the world as a consequence of globalization at the beginning of the new millennium.

On each occasion it was nice to meet Haruko Sanada in Europe (oaidekite totemo-tanoshika ttadesu), and, now it is nice to read her book. The exposition is clear, well arranged into an introduction followed by seven chapters with a short summary at the beginning of each and with a plenty of tables and figures. It is supplied with a brilliant preface by G. Altmann. A list of special terms with explanations, a list of abbreviations, a list of tables and figures, as well as a careful bibliography, a general index, and an index of Japanese words together with their original written forms are benefits for the comfort of the reader.