QUANTITATIVE LINGUISTICS

Volume 53

Editors:
Reinhard Köhler, Burghard Rieger

Editorial Board:
Altmann
Arapov
Boroda
Boy
Brainerd
Embleton
Grotjahn
Köster
Piotrowski
Sambor
Tanaka

G. Altmann (ed.)

Glottometrica 14

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Glottometrica ... -
WVT Wissenschaftlicher Verlag Trier.
14 (1993)
(Quantitative linguistics; Vol. 53)
ISBN 3-88476-081-5
NE: GT

Umschlaggestaltung
Brigitta Disseldorf
smart graphik & design, Trier

© WVT Wissenschaftlicher Verlag Trier
ISBN 3-88476-081-5

Alle Rechte vorbehalten
Nachdruck oder Vervielfältigung nur mit
ausdrücklicher Genehmigung des Verlags
Trier, 1993

WVT Wissenschaftlicher Verlag Trier
BergstraBe 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504

Contents

Glottometrika 14, 1993, 1-53

Graphemes and Context: Statistical data on the graphology of standard Finnish

Matti Pääkkönen, Oulu

1. Material

The principal material for this survey consists of the corpus of current standard Finnish based on statistical sampling and compiled at the Department of Finnish and Lappish, University of Oulu, in 1968 - 1970. It contains material from all the mass media except for television, and a set of reference material consisting of free speech. The main sources for the corpus comprise

1) Finnish language newspapers and magazines from 1967,
2) original Finnish literary and non-fictional writing from 1961 - 1967,
3) radio programmes produced over the period 29.9.1968 - 26.5.1969, and
4) 15 hours of free speech in standard Finnish recorded in 1968 and transcribed for the purposes of the corpus.

No translated text is included. This initial material was then classified into 58 hypothetical context types, each represented by 100 randomly selected passages of five sentences and at least 60 words, i.e. by a total material of 500 sentences and at least 6,000 graphological words. The whole material contains 425,332 graphological words with a total of 3,243,152 graphemes. After the exclusion of any special signs and symbols, the material comprised 421,794 graphological words of a mean length of 7.42 graphemes. The calculations for this part of the material are thus based on a total of 3,130,382 graphemes. Classification and random sampling was carried out according to the following principles[1]:

[1] The material is presented in more detail in the publication "Oulu Corpus. Material for research into standard Finnish in the 1960's. Reports from the Department of Finnish and Lappish at the University of Oulu 1. 1982." Compiled by Pauli Saukkonen.

1) 30 % of both the 222 newspapers and 969 magazines occurring in the various lists available of these were chosen at random for classification. Every article in each newspaper and magazine was classified (e.g. NEWS, CAUSERIES, ADVERTISEMENTS), followed by random selection of the texts and the points at which the samples should begin. The samples in this category contain a total of 106 170 graphological words and 804,842 graphemes.

2) Literary and non-fictional writing was classified according to the UDK system used in scientific libraries, and the works and samples to be included were again decided at random. The samples of literary writing contain a total of 46,934 graphological words or 300,093 graphemes, the figures for non-fictional writing being 177,695 and 1,386,273 respectively.

3) The samples of radio language were selected from spoken programmes originally produced for the radio, e.g. NEWS, RADIO PLAYS, FRAME STORIES. One sample of a certain length was selected from each classified programme in temporal order until the class was full. The Finnish Broadcasting Company edited the sample tapes on the basis of random starting points. The samples were transcribed by students according to the instructions given, using a rough transcription. This means that sandhi assimilations and post-aspiration were sometimes recorded, but mainly left out. The sample of transcribed radio language contains 78,964 graphological words (567,944 graphemes).

4) The last class of texts, free speech, serves mainly as reference material. The speakers of this class of language were chosen by selecting 14 persons living in Helsinki at random from "Who's Who in Finland" and 16 persons who have passed the matriculation examination and are living in Oulu from the population files of the Oulu Police Station. Each subject was interviewed for half an hour, and the beginning of the sample to be taken was selected from the counter on the tape recorder. The samples were transcribed according to the same principles as those of radio language. It would seem that sandhi assimilation and post-aspiration were marked more often in these texts than in the radio language transcriptions, but still not regularly. The number of graphological words in the sample thus totals 12,031 (71,230 graphemes).

The whole of this "Oulu corpus" was duplicated in 30 copies, punched onto cards for computer processing and stored on magnetic tape. The material also exists in the form of running text on microfiche and as Key Word in Context concordances arranged in normal alphabetical order and reverse form (alphabetical order of their last letters).

The material was classified into the following hypothetical sub-classes:

Code no.	Name	Abbreviation	No. of grapho-logical words
	Magazines	MAG	
01	Factual articles	MAG FAC	6,663
02	News	MAG NEW	6,602
03	Causeries	MAG CAU	6,142
04	Devotional writings	MAG DEV	5,892
05	Factual articles for young people	MAG YOU	5,538
06	Reviews on non-fictional literature	MAG NON	7,491
07	Art criticism	MAG ART	7,503
08	Reports	MAG REP	6,703
	Newspapers	NEW	
11	Factual articles	NEW FAC	6,449
12	News	NEW NEW	6,573
13	Reports	NEW REP	6,815
14	Causeries	NEW CAU	7,307
15	Art criticism	NEW ART	7,489
16	Sports coverage	NEW SPR	6,355
	Magazines and newspapers		
22	Advertisements	ADS	6,307
23	Captions	CAP	6,341
	Fiction	FIC	
30	Plays	PLY	6,367
31	Plays for young readers	YPLY	6,358
32	Poetry	POE	6,770
33	Causeries	CAU	7,119
34	Narrative literature	NAR	6,973
35	Narr. literature for young readers	YNAR	6,704
36	Narr. literature	CNAR	6,643
	Non-fiction	NON	
39	Devotional books	NON DEV	7,101
40	Memoirs	(NON)MEM	7,402
41	Travel books	(NON)TRA	6,955
42	Encyclopedias	NON ENC	6,732
43	Laws	(NON)LAW	6,737
44	Religion	(NON)REL	7,877
45	Philosophy	(NON)PHL	8,101

Code no.	Name	Abbreviation	No. of grapho-logical words
46	History	(NON)HST	7,732
47	Social sciences	(NON)SOC	8,139
48	Economics	(NON)ECO	7,631
49	Administration	(NON)ADM	8,054
50	Jurisprudence	(NON)JUR	11,860
51	Committee reports	(NON)COM	9,274
52	Mathematics	(NON)MAT	6,905
53	Biology	(NON)BIO	7,154
54	Medicine	(NON)MED	7,265
55	Agriculture	(NON)AGR	7,030
56	Technology	(NON)TEC	7,165
57	Household management and hobbies	(NON)HOU	7,180
58	Humanistic research	(NON)HUM	7,894
59	Research in social sciences	(NON)SOC RES	8,289
60	Mathematical-technical research	(NON)MAT RES	7,256
61	Biological-medical research	(NON)BIO RES	7,962
	Radio	RAD	
70	News	RAD NEW	6,987
71	Current affairs	RAD CUR	7,146
72	Sports coverage	RAD SPR	6,987
73	Lectures	RAD LEC	7,468
74	Radio plays	RAD PLY	6,484
75	Devotional programmes	RAD DEV	6,809
77	Frame stories for schools	RAD SCH	6,659
78	Frame stories for the young	RAD YNG	6,803
79	Frame stories	RAD FRM	6,888
80	Interviewees' contributions	RAD INT	8,800
81	Discussions	RAD DIS	7,933
90	Free standard speech	FREE	12,031

A second corpus used for the present purpose consists of samples of the language used by Finnish-speaking members of parliament as compiled by Esko Vierikko for his licentiate thesis, and referred to by him as parliamentary language. This sample can be divided into two stylistic categories. *Class 91* represents free speech by the members of parliament, for which purpose Vierikko interviewed 54 members in 1968 - 1969, each interview lasting an hour. Five sections with approx. 200 graphological words were selected from the recordings at regular intervals, i.e. more than 1000 graphological words from each. The speech was transcribed roughly, but in such a manner that sandhi assimilations and most post-aspirations were also included. This class 91 contains 56,512 graphological words and 364,012 graphemes (mean length of a graphological word 6.44 graphemes). *Class 92* consists of official parliamentary language, mainly speeches made in plenary sessions. 2/3 of this was transcribed from the tapes and 1/3 drawn at random from the minutes of parliament and normalized by the secretaries. The sample is almost as large as the previous one, containing 56,094 graphological words and 440,759 graphemes (mean length 7.86 graphemes).

The starting point for the present investigation was a count of the graphemes by text classes. The computer languages used for programming were DOS-AS-SEMBLER and FORTRAN IV. The graphemes were calculated using an IBM S/360 main-frame computer owned by the Computer Centre of Finland and situated in Oulu, the computers used for the other calculations and tests being a Honeywell H 1642 and a Sperry Univac 1100/20 at the University of Oulu and the Univac 1108 of the Finnish National Fund for Research and Development in Helsinki. Graphemes x and z had to be left out since they were used in data acquisition to represent mathematical expressions and unidentified names or words in speech, nor were special characters or groups of characters (e.g. &, $, § and %) processed. The numbers of punctuation marks and figures were processed but not tabulated. As far as punctuation marks are concerned, the full stop and comma are of equal frequency in written language, i.e. approx. 1 % of the total number of characters, punctuation marks and figures, which is higher than that of the graphemes d and ö in Finnish, those which have the lowest frequency. Since the Finnish alphabet does not have a character for /ŋ/, the frequency of g includes both /g/ and /ŋ/, and g mostly represents the sound /ŋ/ (see Kaisa Häkkinen 1977: 63; 1978: 14 - 15).

The present material consists of a total of 534,400 graphological words, containing 3,935,153 graphemes, or characters, the mean number of graphemes per word being 7.36. Just under two thirds of this material represents written language, i.e. 330,799 graphological words containing 2,491,208 graphemes (61.90 % of all

words and 63.31 % of all graphemes), the mean number of graphemes per word being 7.53, just over a third being transcribed speech, represented by 203,601 words and 1,443,945 graphemes (38.10 % of all words and 36.69 % of all graphemes), giving a mean word length of 7.09 graphemes. As described above, the transcriptions of speech usually follow the orthographical system of the written language. The radio language was mainly transcribed as written speech, and the changes caused by sandhi assimilation and post-aspiration were most regularly marked in the transcriptions of interviews with members of parliament. The total number of graphemes is represented by n^1, spoken language by n^2 and written language (n^1 - n^2) by n^3. The frequencies and percentage distributions of graphemes in these three groups are as follows:

TABLE 1. Frequency of graphemes in the whole material (n^1), in transcribed spoken language (n^2) and in written language (n^3).

	n^1			n^2			n^3		
		f	%		f	%		f	%
1.	a	457,350	11.62	a	160,812	11.14	a	296,538	11.90
2.	i	421,366	10.71	i	156,359	10.83	i	265,007	10.64
3.	t	388,711	9.88	t	145,442	10.07	t	243,269	9.77
4.	n	341,181	8.67	n	125,270	8.68	n	215,911	8.67
5.	e	323,087	8.21	e	118,642	8.22	e	204,445	8.21
6.	s	309,350	7.86	s	113,675	7.87	s	195,675	7.85
7.	l	226,627	5.76	l	85,074	5.89	l	141,553	5.68
8.	o	208,923	5.31	o	78,378	5.43	k	132,990	5.34
9.	k	207,520	5.27	ä	74,880	5.19	o	130,545	5.24
10.	u	196,678	5.00	k	74,530	5.16	u	126,164	5.06
11.	ä	189,134	4.81	u	70,514	4.88	ä	114,254	4.59
12.	m	137,972	3.51	m	55,700	3.86	m	82,272	3.30
13.	v	96,316	2.45	v	33,536	2.32	v	62,780	2.52
14.	r	85,116	2.16	j	28,370	1.96	r	57,822	2.32
15.	j	75,961	1.93	r	27,294	1.89	j	47,591	1.91
16.	h	71,733	1.82	y	26,648	1.85	h	45,503	1.83
17.	y	71,316	1.81	h	26,230	1.82	y	44,668	1.79
18.	p	65,358	1.66	p	22,076	1.53	p	43,282	1.74
19.	d	33,148	0.84	d	12,078	0.84	d	21,070	0.85
20.	ö	18,655	0.47	ö	6,467	0.45	ö	12,188	0.49
21.	g	4,151	0.11	g	1,005	0.07	g	3,146	0.13
22.	b	2,068	0.05	b	475	0.03	b	1,593	0.06
23.	f	1,934	0.05	f	395	0.03	f	1,539	0.06
24.	c	1,091	0.03	c	52	0.00	c	1,041	0.04
25.	w	329	0.01	w	22	0.00	w	307	0.01
26.	å	52	0.00	å	20	0.00	å	30	0.00
27.	q	26	0.00	q	1	0.00	q	25	0.00
		3,935,153	100.00		1,443,945	100.00		2,491,208	100.00

As indicated by the Table, the graphemes occur almost in the same order in classes n^1 and n^3: only o and k are reversed in position. The distribution of graphemes in the transcribed spoken language differs more markedly from that in the written language. The significances of the differences between the percentages are presented in Table 4.

804,771 graphemes of parliamentary language, i.e. over a quarter of that material (25.71 %), were added here to the set of graphemes serving as a basis for statistics published earlier. This addition did not cause any significant changes in the proportions established in the earlier material, however. The total number of vowels in the material (n^1) is 1,886,561 (47.94 %) and that of consonants 2,048,592 (52.06 %). The percentages are exactly the same as earlier. The number of consonants per 100 vowels is 108.59, the figure being one hundredth larger than before. The mutual order of frequency of the vowels in this material is as follows:

TABLE 2. Frequency of vowels in the whole material (n^1).

	f	% of all	% of vowels
a	457,350	11.62	24.24
i	421,366	10.71	22.34
e	323,087	8.21	17.13
o	208,923	5.31	11.07
u	196,678	5.00	10.43
ä	189,134	4.81	10.03
y	71,316	1.81	3.78
ö	18,655	0.47	0.99
å	52	0.00	0.00
	1,886,561	47.94	100.00

The sum of the occurrences of the five front vowels (i, e, ä, y, ö), i.e. 1,023,558, is little over a half (54.26 %) of the sum of all vowels, while the number of back vowels (a, o, å, u) is 863,003 (45.74 %). The number of front vowels per 100 back vowels is 118.60 in connected stretches of text. The vowels e and i, which are indifferent in terms of vowel harmony, occur relatively frequently, their total occurrences being 744,453, i.e. 39.46 % of the vowels (18.92 % of all graphemes). The distributions of vowels grouped according to other features are as follows:

- closed (i, y, u)	689,360	(36.54 %)
- semi-closed (e, o, å, ö)	550,717	(29.19 %)
- open (a, ä)	646,484	(34.27 %)
- unrounded (i, e, a, ä)	1,390,937	(73.73 %)
- round (y, ö, u, o, å)	495,624	(26.27 %)

The total number of consonants in the material is 2,048,592 (52.06 %), being represented in the Table below in order of frequency.

TABLE 3. Frequency of consonants in the whole material (n^1).

	f	% of all	% of consonants
t	388,711	9.88	18.97
n	341,181	8.67	16.65
s	309,350	7.86	15.10
l	226,627	5.76	11.06
k	207,520	5.27	10.13
m	137,972	3.51	6.73
v	96,316	2.45	4.70
r	85,116	2.16	4.15
j	75,961	1.93	3.71
h	71,733	1.82	3.50
p	65,358	1.66	3.19
d	33,148	0.84	1.62
g	4,151	0.11	0.20
b	2,068	0.05	0.10
f	1,934	0.05	0.09
c	1,091	0.03	0.05
w	329	0.01	0.02
q	26	0.00	0.00
	2,048,592	52.06	100.00

When examining the frequency of the most significant phonetic features of consonants, c, w, and q, i.e. those which belong mainly to foreign words, can be excluded. The group with the highest frequency is that of dentals (t, d, s, r, l, n), the total number of occurrences being 1,384,133, i.e. 67.57 % of all consonants. The number of labials (p, b, m, f, v) is 303,648 (14.82 %) and that of palato-velar consonants (k, g, j) almost the same, i.e. 287,632 (14.04 %). A computational

inaccuracy should be noted here, in that most of the 4,151 occurrences of the grapheme g apparently represent the velar nasal /ŋ/, but this does not affect the proportions of the graphemes to any significant extent. Voiceless plosives account for almost a third of all consonants (661,589 or 32.29 %), being nevertheless highly different in their frequency. The small group of semi-vowels, i.e. v + j also deserves mention with their 172,277 occurrences, accounting for 8.41 % of all consonant graphemes.

2. Written and spoken language and mean value

The differences between the percentages of occurrence in the transcriptions of written and spoken language and the graphemes in the whole material are relatively large. The significances of the differences were tested by means of the *t*-test and by calculating confidence limits for the percentages. The *t*-test was calculated by means of the formula

$$t = \frac{(K - \mu)\sqrt{N-1}}{s}$$

and χ^2 from the formula $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{n} \frac{O^2_i}{E_i} - N$

The graphemes of these three major groups are presented in the Table below in descending order of percentage frequency of occurrence. Three crosses in the column for the t-test indicate a statistically highly significant difference (at a risk level of 0.1 %), two crosses indicating a significant one (risk 1 %), and one cross an almost significant one (risk 5 %). In the Table below, degrees of significance of the differences between the percentages on the first and second lines are marked on the topmost line of the t-test column for each grapheme, those between the second and third on the middle line and those between the first and third on the last line. The confidence intervals were calculated at a risk of 0.1 %. When testing the significance of differences by means of confidence values, the upper limit of the smaller percentage (CU) is subtracted from the lower limit of the larger percentage (CL). If the difference is positive, it is significant at a probability of 99.9 %. As above, group n^1 contains all the graphemes in the present material (3,935,153 characters), group n^2 those in transcribed speech (1,443,945 characters) and n^3 those in written language (2,491,208 characters).

TABLE 4. Graphemes of the whole material (n^1), spoken language (n^2) and written language (n^3) arranged in descending order of frequency.

		f	%	t-test	CU	CL
a	n^3	296,538	11.90	xxx	11.97	11.83
	n^1	457,350	11.62	xxx	11.67	11.57
	n^2	160,812	11.14	xxx	11.23	11.05
i	n^2	156,359	10.83	xxx	10.92	10.74
	n^1	421,366	10.71	xx	10.76	10.66
	n^3	265,007	10.64	xxx	10.70	10.58
t	n^2	145,422	10.07	xxx	10.15	9.99
	n^1	388,711	9.88	xxx	9.93	9.83
	n^3	243,269	9.77	xxx	9.83	9.71
n	n^2	125,270	8.68		8.76	8.60
	n^1	341,181	8.67		8.72	8.62
	n^3	215,911	8.67		8.73	8.61
e	n^2	118,642	8.22		8.30	8.14
	n^1	323,087	8.21		8.26	8.16
	n^3	204,445	8.21		8.27	8.15
s	n^2	113,675	7.87		7.94	7.80
	n^1	309,350	7.86		7.90	7.82
	n^3	195,675	7.85		7.91	7.79
l	n^2	85,074	5.89	xxx	5.95	5.83
	n^1	226,627	5.76	xxx	5.80	5.72
	n^3	141,553	5.68	xxx	5.73	5.63
o	n^2	78,378	5.43	xxx	5.49	5.37
	n^1	208,923	5.31	xxx	5.35	5.27
	n^3	130,545	5.24	xxx	5.29	5.19
k	n^3	132,990	5.34	xxx	5.39	5.29
	n^1	207,520	5.27	xxx	5.31	5.23
	n^2	74,530	5.16	xxx	5.22	5.10
u	n^3	126,164	5.06	xxx	5.11	5.01
	n^1	196,678	5.00	xxx	5.04	4.96
	n^2	70,514	4.88	xxx	4.94	4.82
ä	n^2	74,880	5.19	xxx	5.25	5.13
	n^1	189,134	4.81	xxx	4.85	4.77
	n^3	114,254	4.59	xxx	4.63	4.55
m	n^2	55,700	3.86	xxx	3.91	3.81
	n^1	137,972	3.51	xxx	3.54	3.48
	n^3	82,272	3.30	xxx	3.34	3.26
v	n^3	62,780	2.52	xxx	2.55	2.49
	n^1	96,316	2.45	xxx	2.48	2.42
	n^2	33,536	2.32	xxx	2.36	2.28
r	n^3	57,822	2.32	xxx	2.35	2.29
	n^1	85,116	2.16	xxx	2.18	2.14
	n^2	27,294	1.89	xxx	1.93	1.85

		f	%	t-test	CU	CL
j	n^2	28,370	1.96	x	2.00	1.92
	n^1	75,961	1.93		1.95	1.91
	n^3	47,591	1.91	xxx	1.94	1.88
h	n^3	45,503	1.83		1.86	1.80
	n^1	71,733	1.82		1.84	1.80
	n^2	26,230	1.82		1.86	1.78
y	n^2	26,648	1.85	xx	1.89	1.81
	n^1	71,316	1.81		1.83	1.79
	n^3	44,668	1.79	xxx	1.82	1.76
p	n^3	43,282	1.74	xxx	1.77	1.71
	n^1	65,358	1.66	xxx	1.68	1.64
	n^2	22,076	1.53	xxx	1.56	1.50
d	n^3	21,070	0.85		0.87	0.83
	n^1	33,148	0.84		0.86	0.82
	n^2	12,078	0.84		0.87	0.81
ö	n^3	12,188	0.49	xxx	0.50	0.48
	n^1	18,655	0.47	xx	0.48	0.46
	n^2	6,467	0.45	xxx	0.47	0.43
g	n^3	3,146	0.13	xxx	0.14	0.12
	n^1	4,151	0.11	xxx	0.12	0.10
	n^2	1,005	0.07	xxx	0.08	0.06
b	n^3	1,593	0.06	xxx	0.07	0.05
	n^1	2,068	0.05	xxx	0.05	0.05
	n^2	475	0.03	xxx	0.03	0.03
f	n^3	1,539	0.06	xxx	0.07	0.05
	n^1	1,934	0.05	xxx	0.05	0.05
	n^2	395	0.03	xxx	0.03	0.03
c	n^3	1,041	0.04	xxx	0.04	0.04
	n^1	1,091	0.03	xxx	0.03	0.03
	n^2	52	0.00	xxx	0.00	0.00
w	n^1	329	0.01		0.01	0.01
	n^3	307	0.01	xxx	0.01	0.01
	n^2	22	0.00	xxx	0.00	0.00
å	n^1	52	0.00		0.00	0.00
	n^3	30	0.00		0.00	0.00
	n^2	20	0.00		0.00	0.00
q	n^1	26	0.00		0.00	0.00
	n^3	25	0.00		0.00	0.00
	n^2	1	0.00		0.00	0.00

The spoken and written languages thus differ to a significant extent at various points at the graphemic level. Among the vowels, a occurs more frequently in written language than in spoken language, its t distribution being as much as 22.67, while even 3.29 corresponds to a 99.9 % level of reliability when the number of degrees of freedom is infinite. The frequency of occurrence of u and of the youngest vowel ö is highly significantly greater in the written language than in the spoken language (t values 7.90 and 5.56), while the vowels i (t value 5.87), o (8.10) and especially ä (26.81) occur much more frequently in the spoken language. In the case of consonants, the great differences in the occurrences of l, m, v, r, and p in particular are of special interest. The higher frequency of the last three of these in the written language is highly significant (t values in the same order 12.38, 28.27 and 15.70), while l and m occur very frequently in spoken language (t values 8.62 and 29.11). Graphemes occurring in specialized loanwords and citations are naturally more frequent in the written language. All in all, statistically highly significant differences occur within the two major groups, spoken and written language, in the case of as many as 20 graphemes. A highly significant difference in frequency of occurrence also is found in 15 out of the 19 graphemes which form the "minimal system" (see F. Karlsson 1983: 65 - 66) of the phoneme paradigm of the Finnish language.

3. Comparison of statistics

In his investigation "Dynamics of the Finnish language" (Suomi 116:3), Vilho Setälä published statistics regarding graphological letters in the language of the New Testament and a table of weighted mean values based on these and earlier statistics, while the corpus used by Kaisa Häkkinen for her relatively extensive statistics of phonemes consists of a Finnish folktale (Häkkinen 1978). The graphemes (or phonemes) of the six largest sets of statistics are presented side by side in the table below in order of frequency. n^1 = mean frequency in the present material (3,935,153 graphemes), n^2 = transcribed speech in the present material (1,443,945 graphemes), n^3 = written language in the present material (2,491,208 graphemes), NT = graphemes in the New Testament (804,243 graphemes), WM = graphemes arranged in the order of weighted mean value as presented by Setälä (1,082,417 graphemes; Setälä 1972: 9 - 12); KH = phonemes in Kaisa Häkkinen's material (166,000 symbols; Häkkinen 1978: 13, 45).

As can be seen, differences abound, and judging from the percentages they are quite considerable differences. The same position is occupied in all sources only by a (1.), i (2.), e (5.), s (6.) and m (12.). The order of frequency is the same up

to the seventh grapheme, which is l, in all three main groups of the present material and in Setälä's table of weighted mean values, while n has risen to 3rd position among the graphemes of the NT and in the set of phonemes in Kaisa Häkkinen's material, seventh position being occupied by ä in the language of the NT and by k in KH. As far as the other positions are concerned, only the 12th position is the same for all sources, being occupied by m.

TABLE 5. Order of frequency of graphemes in the present material (n^1), the transcribed spoken language in the present material (n^2), the written language in the present material (n^3), Setälä's material from the New Testament (NT), the table of weighted material (KH).

	n^1	n^2	n^3	NT	WM	KH
1.	a	a	a	a	a	a
2.	i	i	i	i	i	i
3.	t	t	t	n	t	n
4.	n	n	n	t	n	t
5.	e	e	e	e	e	e
6.	s	s	s	s	s	s
7.	l	l	l	ä	l	k
8.	o	o	k	l	ä	l
9.	k	ä	o	k	k	ä
10.	u	k	u	o	o	u
11.	ä	u	ä	u	u	o
12.	m	m	m	m	m	m
13.	v	v	v	h	v	p
14.	r	j	r	j	h	v
15.	j	r	j	v	j	r
16.	h	y	h	p	r	j
17.	y	h	y	r	p	h
18.	p	p	p	y	y	y
19.	d	d	d	d	d	ö
20.	ö	ö	ö	ö	ö	d
21.	g	g	g	g	f	ŋ
22.	b	b	b	f	g	-
23.	f	f	f	b	b	-
24.	c	c	c	c	c	-
25.	w	w	w	-	-	-
26.	å	å	å	-	-	
27.	q	q	q	-	-	-

TABLE 6. Frequencies of graphemes (%) in the present material (n^1), Setälä's material from the New Testament (NT), the table of weighted mean values compiled by Setälä (WM) and Kaisa Häkkinen's material (KH).

		n^1	NT	t^1	WM	t^2	KH	t^3
1.	a	11.62	11.93	xxx	12.20	xxx	12.34	xxx
2.	i	10.71	9.95	xxx	10.80	xx	11.42	xxx
3.	t	9.88	9.28	xxx	9.90		8.79	xxx
4.	n	8.67	9.83	xxx	9.60	xxx	9.83	xxx
5.	e	8.21	9.31	xxx	8.90	xxx	7.35	xxx
6.	s	7.86	7.30	xxx	7.10	xxx	7.16	xxx
7.	l	5.76	5.70	x	5.70	x	5.78	
8.	o	5.31	4.66	xxx	5.00	xxx	4.48	xxx
9.	k	5.27	5.09	xxx	5.20	xx	6.15	xxx
10.	u	5.00	4.69	xxx	4.80	xxx	4.88	x
11.	ä	4.81	5.70	xxx	5.60	xxx	5.61	xxx
12.	m	3.51	3.29	xxx	3.50		2.81	xxx
13.	v	2.45	2.20	xxx	2.30	xxx	2.19	xxx
14.	r	2.16	1.59	xxx	1.70	xxx	2.15	
15.	j	1.93	2.72	xxx	1.90	x	2.14	xxx
16.	h	1.82	2.62	xxx	1.90	xxx	2.10	xxx
17.	y	1.81	1.18	xxx	1.40	xxx	1.59	xxx
18.	p	1.66	1.52	xxx	1.60	xxx	2.30	xxx
19.	d	0.84	0.83		0.80	xxx	0.40	xxx
20.	ö	0.47	0.29	xxx	0.40	xxx	0.43	x
21.	g	0.11	0.10	x	0.09	xxx	-	-
22.	b	0.05	0.06	xxx	0.01	xxx	-	-
23.	f	0.05	0.06	xxx	0.04	xxx	-	-
24.	c	0.03	-	-	0.00	xxx	-	-
25.	w	0.01	-	-	-	-	-	-
26.	å	0.00	-	-	-	-	-	-
27.	q	0.00	-	-	-	-	-	-

Table 6 is a comparison of the mean percentages in the present material (n^1), the percentages for the New Testament presented by Setälä (NT) and his weighted mean values (WM) and the percentages in Häkkinen's material (KH). The significance of the differences between the percentages is tested by means of the t-test. t^1 = degree of significance between n^1 and NT, t^2 = degree of significance between n^1 and PK, and t^3 = degree of significance between n^1 and KH.

All four sets of statistics are based on a relatively extensive source material. Comparison of the language of the NT with n^1 indicates that only l, d and g are represented almost in the same proportions. Highly significant differences occur in the frequencies of all the other graphemes. The weighted mean values for t, l, m and j do not differ from the means in the present material to any significant extent, but the differences between the frequencies of all the other graphemes are significant or highly significant. It should be noted that the set of graphemes used in the NT was the dominant factor in the calculation of the weighted mean value. What can the reason be for such great differences? Setälä (p. 10) also calculated the frequencies of certain words in the NT, noting that the words *ja* 'and', *Jeesus* 'Jesus', *Jumala* 'God', *juutalaiset* 'the Jews' occurred altogether 12,000 times and account for 10 % of the total vocabulary, and likewise *he* 'they', *henki* 'spirit', *herra* 'Lord', *hän* 'he', *ihminen* 'man', *ihme* 'miracle' and *hyvä* 'good' occurred altogether 10,500 times (8.5 % of the vocabulary). Since j and h occur considerably more frequently in the NT than in the normal language n^1 (t values 45.41 and 47.21), he has weighted them less than the other graphemes in the NT when calculating the mean values. Despite this the weighted mean value for h differs to a significant extent from the corresponding value in n^1. But Setälä did not pay attention to the extraordinary dominance of n and e, for instance, and especially that of ä (t values 33.35, 32.41, and 33.49) in the language of the NT, nor did he consider the obvious under-representation of r, y and ö, for instance (t values 32.77, 39.79 and 22.24 as compared with n^1). The vocabulary could also be examined for the causes of the high frequency of ä (*tänä päivänä* 'today', *täyttää* 'fulfil', *säätää* 'decree', *päästää* 'free', *kärsiä* 'suffer' etc.; according to Setälä (p. 43) the frequency of the pair ää is 5.658 % in the NT), but accurate comparisons cannot be carried out without a frequency dictionary for the NT. In any case, the frequencies of the graphemes in the NT differ to an appreciable extent from the mean values in the present material.

As compared with the statistics compiled by Kaisa Häkkinen, no significant differences occur in the percentages of l, u, r, and ö, whereas in the case of all the other graphemes the difference is highly significant. The representation of n (t'= 16.41), k (15.67), ä (14.87) and p (19.84) is considerably above the average

in Häkkinen's material, while o (14.82), m (15.24) and d (19.45) are considerably below it. Speculation about the reasons for this is not relevant here, but it should be noted that the source of her material, "Tarulinna" by Lehtonen, uses d at least in the manner typical of the standard language even in its representations of speech.

Anneli Pajunen and Ulla Palomäki have examined the frequency distribution of phonetic symbols in texts from the Finnish Syntactic Archives transcribed by means of the rough system used for the Finnish dialects (1984: 73 - 89). What particularly attracts attention in their statistics is that the vowel i occupies the first position in all of the dialects, its proportion varying between 11.4 % in the Savo dialect and 13.8 % in the southeastern dialects. The position of a, the most frequently occurring vowel in all other statistics, is as low as 4th in the south-western dialects (8.8 %), 3rd in the Savo dialects, and 2nd in the Häme dialects (9.7 %, 10.0 %, and 10.2 % respectively). Pajunen and Palomäki would appear to be on the right track in their search for an explanation for this phenomenon in suggesting that the increase in the frequency of i in the dialects and the decrease in a is connected with the structure of the later syllables in words. Their investigations indicate that 50 % of all graphological words in the written language have an inflectional morph attached to them, for instance, but only 30 % of those in the spoken language, and it is true that the majority of the nominal case endings terminate in A. Also, some vowel combinations ending in A in the later syllables of words have become contracted to long vowels (e.g. *korkee*, *pimee*, *kaikkii*, *pappii*, *maitoo*, *tyttöö*) (Pajunen & Palomäki 1984: 48 - 59, 73 - 89). Olli Järvikoski, having counted graphemes and phonemes partly in the same material as Pajunen and Palomäki, suggests that the frequency of i is based on apocope. According to his calculations, the proportion of a in word final position is 16.8 % in dialects and 25.2 % in the written language, while the relative frequency of final i is 13.9 % in dialects and 10.3 % in the written language (1985: 42). It should also be mentioned here that according to Pauli Saukkonen (1977: IX), a accounts for as much as 27 % of the letters occurring in final positions in graphological word forms in modern Finnish, but i for only 8 %.

The reasons for the distinct differences between the sets of graphemes and phonemes in the above major groups will not be discussed any further here, but instead we shall now go on to examine the somewhat smaller main classes, which represent purely written language and which have also been studied in connection with other investigations into this standard language material, i.e. the language of the press, fiction, and non-fiction. Statistical data on transcribed radio language, free standard spoken language and parliamentary language will

also be presented for the sake of comparison. Phonemes in the final position in graphological words in the language of these main classes were transcribed so inaccurately that they cannot be reliably compared either with written language or with an accurately transcribed dialect.

4. Newspaper language (style classes 01 - 23)

The present sample of newspaper language contains 106,170 graphological words and 804,842 graphemes, the mean number of graphemes per word being 7.58. The graphemes are presented in the table below in order of frequency, E^1 being here, as in the other tables of graphemes of these main classes, the deviation of the frequency of each grapheme in the main class from its frequency in n^1 expressed in percentage points. The significances of the differences were tested by means of the t-test. The column (n^1) presents the graphemes in their average order of frequency.

The mutual order of the graphemes in newspaper language is very close to the average. The only graphemes which have changed their positions are p and y, the former having risen to 17th position. The occurrence of a is highly significantly above the average in newspaper language and the figures for ä and y highly significantly below the average, the t values being 12.21, 15.77, and 6.16 respectively. The frequencies of the other vowels do not involve any differences. The frequencies of the consonants r, j, h, p, g, b, f, c, and w are highly significantly greater in newspaper language, and correspondingly, the frequencies of t, s, and m, and surprisingly also that of d, are below the average. The frequencies of l and k are almost significantly above the average.

In his study "Sanamuodot ja niiden kirjainrakenne suomenkielisessä sanomalehtitekstissä" ("Word forms and their grapheme structure in Finnish newspaper texts"; 1974), Jaakko Pesonen examined a sample containing 22,211 graphological words and 174,513 graphemes. The mean number of graphemes per word in his material was 7.86 (p. 10). The order of frequency of the graphemes was exactly the same as in the present material, but some highly significant differences arise between these two samples of newspaper language, in that the frequencies of o, ö, and g are significantly higher in Pesonen's material, while those of ä and m are correspondingly lower. The reasons for such differences are difficult to find, but it may be noted that the size of the sample examined by Pesonen was only a quarter of that employed here.

TABLE 7. Graphemes in newspaper language.

	(n^1)	NEW	f	%	E^1	t-test
1.	(a)	a	97,380	12.10	+0.48	xxx
2.	(i)	i	85,609	10.64	-0.07	
3.	(t)	t	77,181	9.59	-0.29	xxx
4.	(n)	n	70,164	8.72	+0.05	
5.	(e)	e	65,927	8.19	-0.02	
6.	(s)	s	61,683	7.66	-0.20	xxx
7.	(l)	l	46,858	5.82	+0.06	x
8.	(o)	o	42,999	5.34	+0.03	
9.	(k)	k	42,864	5.33	+0.06	x
10.	(u)	u	40,166	4.99	-0.01	
11.	(ä)	ä	35,420	4.40	-0.41	xxx
12.	(m)	m	27,202	3.38	-0.13	xxx
13.	(v)	v	19,840	2.47	+0.02	
14.	(r)	r	18,973	2.36	+0.20	xxx
15.	(j)	j	16,329	2.03	+0.10	xxx
16.	(h)	h	15,312	1.90	+0.08	xxx
17.	(y)	p	14,263	1.77	+0.11	xxx
18.	(p)	y	13,786	1.71	-0.10	xxx
19.	(d)	d	6,406	0.80	-0.04	xxx
20.	(ö)	ö	3,665	0.46	-0.01	
21.	(g)	g	1,161	0.14	+0.03	xxx
22.	(b)	b	557	0.07	+0.02	xxx
23.	(f)	f	506	0.06	+0.01	xxx
24.	(c)	c	400	0.05	+0.02	xxx
25.	(w)	w	164	0.02	+0.01	xxx
26.	(å)	å	17	0.00	0.00	
27.	(q)	q	10	0.00	0.00	
			804,842	100.00		

5. Fiction (style classes 30 - 36)

The sample drawn from fiction is fairly small, the number of graphological words being 46,934, containing 300,093 graphemes. The graphological words are very markedly shorter than average, their mean length being only 6.39 graphemes. The graphemes were tabulated as above, i.e. in order of frequency. E^1 = Deviation of the frequencies of the graphemes in FIC from those in n^1 expressed in percentage points.

TABLE 8. Graphemes in fiction.

	(n^1)	FIC	f	%	E^1	t-test
1.	(a)	a	35,221	11.74	+0.12	x
2.	(i)	i	33,898	11.30	+0.59	xxx
3.	(t)	t	28,022	9.34	-0.54	xxx
4.	(n)	n	27,280	9.09	+0.42	xxx
5.	(e)	e	23,370	7.79	-0.42	xxx
6.	(s)	s	20,889	6.96	-0.90	xxx
7.	(l)	l	17,566	5.85	+0.09	x
8.	(o)	k	17,245	5.75	+0.48	xxx
9.	(k)	ä	16,736	5.58	+0.77	xxx
10.	(u)	u	15,547	5.18	+0.18	xxx
11.	(ä)	o	14,311	4.77	-0.54	xxx
12.	(m)	m	9,557	3.18	-0.33	xxx
13.	(v)	v	7,209	2.40	-0.05	
14.	(r)	h	6,931	2.31	+0.49	xxx
15.	(j)	r	6,377	2.13	-0.03	
16.	(h)	p	6,151	2.05	+0.39	xxx
17.	(y)	j	5,686	1.89	-0.04	
18.	(p)	y	4,751	1.58	-0.23	xxx
19.	(d)	d	1,711	0.57	-0.27	xxx
20.	(ö)	ö	1,195	0.40	-0.07	xxx
21.	(g)	g	225	0.07	-0.04	xxx
22.	(b)	b	95	0.03	-0.02	xxx
23.	(f)	f	73	0.02	-0.03	xxx
24.	(c)	c	31	0.01	-0.02	xxx
25.	(w)	w	13	0.00	-0.01	xxx
26.	(å)	q	3	0.00	0.00	
27.	(q)	å	-	-	-	
			300,093	100.00		

The language of fiction thus deviates to a surprising extent from the average, in that as many as nine graphemes occupy a position deviating from the average, although the percentages of y, r, and j are so close to the average that the deviation is not significant in statistical terms. The deviation of a and l is almost significant (5 % risk), and that of the other graphemes highly significant (0.1 % risk). The deviation in the frequencies of s and ä is conspicuous, in that the occurrence of s in fiction is almost a percentage point lower, while that of ä is almost as much higher (t-values 17.73 and 18.90). The higher frequencies of i, n, k, h, and p (t-values 10.06, 7.87, 11.31, 19.18, and 15.99), and correspondingly the lower frequencies of t, e, o, m, v, and d (t-values 9.57, 8.09, 12.76, 9.50, 9.15, and 15.80), also stand out. The differences are thus highly prominent.

6. Language of non-fiction (style classes 39 - 61)

The sample taken from non-fiction forms the largest of the main classes and probably the most heterogeneous as well. It contains 177,695 graphological words and 1,386,273 graphemes, the mean number of graphemes per word being 7.8, i.e. well above the average. The graphemes of this sample are presented in the table below in order of frequency. E^1 = Deviation of the frequencies of the graphemes in NON from those in n^1 expressed in percentage points.

The graphemes of non-fiction account for 35.23 % of the whole material, and only a few deviations in position in the table are to be found. y has risen by two positions, being 15th here, so that j and h have dropped by one place each, nevertheless retaining their mutual order. f and b have also changed positions. As compared with the mean percentage, the deviations in the frequency percentages of o, k, p, w, å, and q are statistically insignificant. The deviation of t is significant and that of the other graphemes highly significant. The highest t-test values are attached to the differences between the frequencies for s (11.24), ä (15.75), m (12.76), r (12.41), h (10.71), and d (9.85), the frequencies of s, r, and d being well above the average, and those of ä, m, and h below it.

TABLE 9. Graphemes in non-fiction.

	(n^1)	NON		%	E^1	t-test
1.	(a)	a	163,937	11.83	+0.21	xxx
2.	(i)	i	145,500	10.50	-0.21	xxx
3.	(t)	t	138,066	9.96	+0.08	xx
4.	(n)	n	118,467	8.55	-0.12	xxx
5.	(e)	e	115,148	8.31	+0.10	xxx
6.	(s)	s	113,103	8.16	+0.30	xxx
7.	(l)	l	77,129	5.56	-0.20	xxx
8.	(o)	o	73235	5.28	-0.03	
9.	(k)	k	72,881	5.26	-0.01	
10.	(u)	u	70,451	5.08	+0.08	xxx
11.	(ä)	ä	62,098	4.48	-0.33	xxx
12.	(m)	m	45,513	3.28	-0.23	xxx
13.	(v)	v	35,731	2.58	+0.13	xxx
14.	(r)	r	32,472	2.34	+0.18	xxx
15.	(j)	y	26,131	1.88	+0.07	xxx
16.	(h)	j	25,576	1.84	-0.09	xxx
17.	(y)	h	23,260	1.68	-0.14	xxx
18.	(p)	p	22,868	1.65	-0.01	
19.	(d)	d	12,953	0.93	+0.09	xxx
20.	(ö)	ö	7,328	0.53	+0.06	xxx
21.	(g)	g	1,760	0.13	+0.02	xxx
22.	(b)	f		0.07	+0.02	xxx
23.	(f)	b		0.07	+0.02	xxx
24.	(c)	c		0.04	+0.01	xxx
25.	(w)	w		0.01	0.00	
26.	(å)	å		0.00	0.00	
27.	(q)	q		0.00	0.00	
			1,386,273	100.00		

7. Comparison of graphemes in newspapers, fiction, non-fiction, and the New Testament

In order to be able to compare the occurrences of graphemes and their variations in the main classes of written language discussed above, the frequencies of all the grapheme are tabulated once more in descending order in tables which include the statistics on the language of the New Testament compiled by Vilho Setälä to represent a special variety of written language. The significance of the differences can be tested by means of the confidence figures in the tables. The figure indicating the upper limit of the confidence interval of the smaller percentage (UL) is subtracted from the lower limit of the confidence interval of the larger percentage (LL). If the difference is positive, it is significant at a risk of 0.1 %.

The grapheme a is clearly the most common in newspapers. Its frequency deviates only almost significantly from the language of the New Testament (risk 5 %), but is highly significantly lower in the classes NON and FIC than in newspaper language (risk 0.1 %). The frequency of a in the language of fiction is also almost significantly lower than in the language of the New Testament, but the other differences are entirely non-significant.

The situation with i is almost the reverse, fiction containing a higher frequency of i than any other main class. The difference between newspaper language and non-fiction is only almost significant, but the frequency of i in the language of the New Testament is to a significant extent lower than in any other class.

The NT is again distinctly different from the other classes in the table for e, in that the number of occurrences of this grapheme is highly significantly greater than in the other main classes, while NON and NEW do not differ significantly and the frequency of e in FIC is highly significantly lower than in any other class.

The frequency of o is highest in newspaper and non-fictional language, but no statistical differences occur in this respect. FIC and NT form a second pair, in which o occurs less than in the former pair to a highly significant extent.

The frequency of u is similar in the categories of FIC, NON and NEW at a risk level of 0.1 %, being highly significantly lower in NT.

TABLE 10. Frequencies of vowels in the languages of newspapers, fiction, non-fiction and the New Testament.

		f	%	CU	CL
a	NEW	97,380	12.10	12.22	11.98
	NT	95,979	11.93	12.05	11.81
	NON	163,937	11.83	11.92	11.74
	FIC	35,221	11.74	11.93	11.55
i	FIC	33,898	11.30	11.49	11.11
	NEW	85,609	10.64	10.75	10.53
	NON	145,500	10.50	10.59	10.41
	NT	80,059	9.95	10.06	9.84
e	NT	74,954	9.31	9.42	9.20
	NON	115,148	8.31	8.39	8.23
	NEW	65,927	8.19	8.29	8.09
	FIC	23,370	7.79	7.95	7.63
o	NEW	42,999	5.34	5.42	5.26
	NON	73,235	5.28	5.34	5.22
	FIC	14,311	4.77	4.90	4.64
	NT	37,483	4.66	4.74	4.58
u	FIC	15,547	5.18	5.31	5.05
	NON	70,451	5.08	5.14	5.02
	NEW	40,166	4.99	5.07	4.91
	NT	37,721	4.69	4.77	4.61
ä	NT	45,867	5.70	5.79	5.61
	FIC	16,736	5.58	5.72	5.44
	NON	62,098	4.48	4.54	4.42
	NEW	35,420	4.40	4.48	4.32
y	NON	26,131	1.81	1.92	1.84
	NEW	13,786	1.71	1.76	1.66
	FIC	4,751	1.58	1.66	1.50
	NT	9,491	1.18	1.22	1.14
ö	NON	7,328	0.53	0.55	0.51
	NEW	3,665	0.46	0.48	0.44
	FIC	1,195	0.40	0.44	0.36
	NT	2,347	0.29	0.31	0.27
å	NEW	17	0.00	0.00	0.00
	NON	13	0.00	0.00	0.00
	FIC	-	-	-	-
	NT	-	-	-	-

The frequency of ä is considerably higher in the New Testament and in fiction than in the other categories, but the difference is not significant, nor is that between newspaper and non-fictional language, although it is indeed large. The difference between the occurrences of ä in NT and NON is thus 1.22 percentage points, the t-test value being as high as 40.21.

The occurrences of y have two separate peaks, NON having highly significantly more ys than the other classes and NT correspondingly less. NEW and FIC also differ to a significant extent (1 % risk).

The vowel ö, which is the youngest vowel in terms of the history of the Finnish language, has its highest frequency in non-fictional language, a feature which is statistically highly significant, and in the same manner as y, it occurs least in the language of the NT. The difference between FIC and NT is highly significant. The occurrence of ö is highly significantly greater (1 % risk) in newspaper language than in fiction. å occurs so little in the samples that differences can be expressed only in tenths of a percent.

Of the successive groups, only NON and NEW differ in the table for t to a highly significant extent. The percentages decrease evenly and the language of NT thus has proportionally the least occurrences.

The mutual order of classes in the table for n is entirely the reverse of the above, the language of NT having a conspicuously high occurrence of n, as much as 1.28 percentage points more than in the language of FIC, which has the lowest. The differences NT - FIC and FIC - NEW are highly significant, while that between NEW and FIC is only significant (1 % risk).

The main class FIC is sometimes the first and sometimes the last in the tables for the four consonants with the highest frequencies, s occurring highly significantly more and l significantly or highly significantly less than in the other classes. Differences between all the main categories in the table for s are statistically highly significant. The occurrences of l are fairly similar in the categories FIC, NEW and NT, with only NON clearly deviating from the others.

The grapheme k has a very high frequency in fiction and a very low one in the language of the New Testament, with no significant differences between newspaper language and non-fiction. As far as the frequencies of the other consonants are concerned, the preponderance of j and h in the language of the New Testament and that of p in literary language may be said to be conspicuous.

TABLE 11. Frequencies of consonants in the language of newspapers, fiction, non-fiction, and the New Testament.

		f	%	CU	CL
t	NON	138,066	9.96	10.04	9.88
	NEW	77,181	9.59	9.70	9.48
	FIC	28,022	9.34	9.52	9.16
	NT	74,684	9.28	9.39	9.17
n	NT	79,114	9.83	9.94	9.72
	FIC	27,280	9.09	9.26	8.92
	NEW	70,164	8.72	8.82	8.62
	NON	118,467	8.55	8.63	8.47
s	NON	113,103	8.16	8.24	8.08
	NEW	61,683	7.66	7.76	7.56
	NT	58,764	7.30	7.40	7.20
	FIC	20,889	6.96	7.11	6.81
l	FIC	17,566	5.85	5.99	5.71
	NEW	46,858	5.82	5.91	5.73
	NT	45,882	5.70	5.79	5.61
	NON	77,129	5.56	5.62	5.50
k	FIC	17,245	5.75	5.89	5.61
	NEW	42,864	5.33	5.41	5.25
	NON	72,881	5.26	5.32	5.20
	NT	40,941	5.09	5.17	5.01
m	NEW	27,202	3.38	3.45	3.31
	NT	26,469	3.29	3.36	3.22
	NON	45,513	3.28	3.33	3.23
	FIC	9,557	3.18	3.29	3.07
v	NON	35,731	2.58	2.62	2.54
	NEW	19,840	2.47	2.53	2.41
	FIC	7,209	2.40	2.49	2.31
	NT	17,751	2.20	2.25	2.15
r	NEW	18,973	2.36	2.42	2.30
	NON	32,472	2.34	2.38	2.30
	FIC	6,377	2.13	2.22	2.04
	NT	12,842	1.59	1.64	1.54
j	NT	21,945	2.72	2.78	2.66
	NEW	16,329	2.03	2.08	1.98
	FIC	5,686	1.89	1.97	1.81
	NON	25,576	1.84	1.88	1.80
h	NT	21,102	2.62	2.68	2.56
	FIC	6,931	2.31	2.40	2.22
	NEW	15,312	1.90	1.95	1.85
	NON	23,260	1.68	1.72	1.64

		f	%	CU	CL
p	FIC	6,151	2.05	2.14	1.96
	NEW	14,263	1.77	1.82	1.72
	NON	22,868	1.65	1.69	1.61
	NT	12,230	1.52	1.57	1.47
d	NON	12,953	0.93	0.90	0.96
	NT	6,715	0.83	0.80	0.86
	NEW	6,406	0.80	0.83	0.77
	FIC	1,711	0.57	0.62	0.52
g	NEW	1,161	0.14	0.15	0.13
	NON	1,760	0.13	0.14	0.12
	NT	834	0.10	0.11	0.09
	FIC	225	0.07	0.09	0.05
b	NON	941	0.07	0.08	0.06
	NEW	557	0.07	0.08	0.06
	NT	530	0.06	0.07	0.05
	FIC	95	0.03	0.04	0.02
f	NON	960	0.07	0.08	0.06
	NT	530	0.06	0.07	0.05
	NEW	506	0.06	0.07	0.05
	FIC	73	0.02	0.03	0.01
c	NEW	400	0.05	0.06	0.04
	NON	610	0.04	0.05	0.03
	FIC	31	0.01	0.02	0.00
	NT	-			
w	NEW	164	0.02	0.03	0.01
	NON	130	0.01	0.01	0.01
	FIC	13	0.00	0.00	0.00
	NT	-			
q	NON	12	0.00	0.00	0.00
	NEW	10	0.00	0.00	0.00
	FIC	3	0.00	0.00	0.00
	NT	-			

The ratio between vowels and consonants in the whole material is 100:108.59. The occurrence of consonants is slightly above average in the written language, the total number of graphemes in style categories 01 - 61 (= n^3) being 2,491,208, of which 1,193,839 (47.92 %) are vowels and 1,297,369 (52.08 %) consonants, i.e. 108.67 consonants per 100 vowels. The corresponding ratio for spoken language (= n^2, style categories 70 - 92) is 100:108.44 (n^2 = 1,443,945, of which 692,722, i.e. 47.97 %, are vowels and 751,223, i.e. 52.03 %, consonants). The differences between the main categories discussed here are even greater.

TABLE 12. Ratio of vowels to consonants in the language of newspapers, fiction, non-fiction, and the New Testament.

Category		Vowels		Consonants		Vow. : Cons.
	n	f	%	f	%	
NEW	804,842	384,969	47.83	419,873	52.17	100 : 109.07
FIC	300,093	145,029	48.33	155,064	51.67	100 : 106.92
NON	1,386,273	663,841	47.89	722,432	52.11	100 : 108.83
NT	804,243	383,910	47.74	420,333	52.26	100 : 109.49

The frequency of consonants per 100 vowels is thus lowest in the fiction material, the differences between the frequencies of vowels and consonants being statistically highly significant as compared with the n^1 material and the language of newspapers, non-fiction and the New Testament. Correspondingly, the figure is highest in the language of the New Testament, the differences as compared with the n^1 material being significant at a 1 % risk. No significant differences exist between the other percentages.

8. Inflectional forms and graphemes

Although the main categories of written language in the present material, i.e. the languages of newspapers, fiction, and non-fiction, are fairly uneven in both size and composition, they can be tentatively compared in terms of the sets of graphemes in the inflectional forms with the highest frequencies. The newspaper language sample has a total of 106,170 inflectional forms with 804,842 graphemes, the mean number of graphemes per word being 7.58, the corresponding figures for fiction being 46,934, 300,093 and 6.39, and for the non-fictional sample 177,695, 1,386,273, and 7.80 respectively. The inflectional forms which occurred at least 100 times were selected from each main category, after which the numbers of graphemes occurring in them were calculated and compared with the total set of graphemes in each main category, the graphemes in the main categories being first tabulated to facilitate comparison. The topmost line in the t-test column indicates the significance of the difference between the first and second lines, the middle one that between the second and third lines and the bottom one that between the first and third lines.

TABLE 13. Graphemes in the language of newspapers, fiction and non-fiction

		f	%	t-test	CU	CL
a	NEW	97,380	12.10	xxx	12.22	11.98
	NON	163,937	11.83		11.92	11.74
	FIC	35,221	11.74	xxx	11.93	11.55
i	FIC	33,898	11.30	xxx	11.49	11.11
	NEW	85,609	10.64	xx	10.75	10.53
	NON	145,500	10.50	xxx	10.59	10.41
t	NON	138,066	9.96	xxx	10.04	9.88
	NEW	77,181	9.59	xxx	9.70	9.48
	FIC	28,022	9.34	xxx	9.52	9.16
n	FIC	27,280	9.09	xxx	9.26	8.92
	NEW	70,164	8.72	xxx	8.82	8.62
	NON	118,467	8.55	xxx	8.63	8.47
e	NON	115,148	8.31	xx	8.39	8.23
	NEW	65,927	8.19	xxx	8.29	8.09
	FIC	23,370	7.79	xxx	7.95	7.63
s	NON	113,103	8.16	xxx	8.24	8.08
	NEW	61,683	7.66	xxx	7.76	7.56
	FIC	20,889	6.96	xxx	7.11	6.81
l	FIC	17,566	5.85		5.99	5.71
	NEW	46,858	5.82	xxx	5.91	5.73
	NON	77,129	5.56	xxx	5.62	5.50
o	NEW	42,999	5.34		5.42	5.26
	NON	73,235	5.28	xxx	5.34	5.22
	FIC	14,311	4.77	xxx	4.90	4.64
k	FIC	17,245	5.75	xxx	5.89	5.61
	NEW	42,864	5.33	x	5.41	5.25
	NON	72,881	5.26	xxx	5.32	5.20
u	FIC	15,547	5.18	x	5.31	5.05
	NON	70,451	5.08	xx	5.14	5.02
	NEW	40,166	4.99	xxx	5.07	4.91
ä	FIC	16,736	5.58	xxx	5.72	5.44
	NON	62,098	4.48	xx	4.54	4.42
	NEW	35,420	4.40	xxx	4.48	4.32
m	NEW	27,202	3.38	xxx	3.45	3.31
	NON	45,513	3.28	xx	3.33	3.23
	FIC	9,557	3.18	xxx	3.29	3.07
v	NON	35,731	2.58	xxx	2.62	2.54
	NEW	19,840	2.47	x	2.53	2.41
	FIC	7,209	2.40	xxx	2.49	2.31

		f	%	t-test	CU	CL
r	NEW	18,973	2.36		2.42	2.30
	NON	32,472	2.34	xxx	2.38	2.30
	FIC	6,377	2.13	xxx	2.22	2.04
j	NEW	16,329	2.03	xxx	2.08	1.98
	FIC	5,686	1.89		1.97	1.81
	NON	25,576	1.84	xxx	1.88	1.80
h	FIC	6,931	2.31	xxx	2.40	2.22
	NEW	15,312	1.90	xxx	1.95	1.85
	NON	23,260	1.68	xxx	1.72	1.64
y	NON	26,131	1.88	xxx	1.92	1.84
	NEW	13,786	1.71	xxx	1.76	1.66
	FIC	4,751	1.58	xxx	1.66	1.50
p	FIC	6,151	2.05	xxx	2.14	1.96
	NEW	14,263	1.77	xxx	1.82	1.72
	NON	22,868	1.65	xxx	1.69	1.61
d	NON	12,953	0.93	xxx	0.96	0.90
	NEW	6,406	0.80	xxx	0.83	0.77
	FIC	1,711	0.57	xxx	0.62	0.52
ö	NON	7,328	0.53	xxx	0.55	0.51
	NEW	3,665	0.46	xxx	0.48	0.44
	FIC	1,195	0.40	xxx	0.44	0.36
g	NEW	1,161	0.14		0.15	0.13
	NON	1,760	0.13	xxx	0.14	0.12
	FIC	225	0.07	xxx	0.09	0.05
b	NON	941	0.07		0.08	0.06
	NEW	557	0.07	xxx	0.08	0.06
	FIC	95	0.03	xxx	0.04	0.02
f	NON	960	0.07	xx	0.08	0.06
	NEW	506	0.06	xxx	0.07	0.05
	FIC	73	0.02	xxx	0.03	0.01
c	NEW	400	0.05	xxx	0.06	0.04
	NON	610	0.04	xxx	0.05	0.03
	FIC	31	0.01	xxx	0.02	0.00
w	NEW	164	0.02	xxx	0.03	0.01
	NON	130	0.01	xxx	0.01	0.01
	FIC	13	0.00	xxx	0.00	0.00
å	NEW	17	0.00	-	-	-
	NON	13	0.00	-	-	-
	FIC	-	-	-	-	-
q	NON	12	0.00	-	-	-
	NEW	10	0.00	-	-	-
	FIC	3	0.00	-	-	-

The differences between the main categories are thus for the most part very clear, since out of the 75 cases in which their significances were tested, 7 were not statistically significant, 3 were almost significant (5 % risk), 6 significant (1 % risk) and 59 highly significant (0.1 % risk).

The following inflectional forms occurred at least 100 times in each main category. The percentages for the occurrences were calculated on the basis of the number of graphological words in each main category.

2 graphemes

NEW	f	%	FIC	f	%	NON	f	%
ja	3,945	3.70	ja	1,874	3.99	ja	6,468	3.64
on	3,032	2.84	on	757	1.61	on	5,169	2.91
ei	856	0.80	ei	519	1.11	ei	1,423	0.80
se	555	0.52	se	488	1.04	se	720	0.41
jo	303	0.28	he	150	0.32	jo	323	0.18
ne	162	0.15	jo	149	0.32	ne	318	0.18
he	100	0.09	ne	128	0.27			
			en	120	0.26			
7	8,953	8.43	8	4,185	8.92	6	14,421	8.12

3 graphemes

NEW	f	%	FIC	f	%	NON	f	%
oli	767	0.72	oli	694	1.48	oli	1,183	0.67
hän	435	0.41	hän	492	1.05	tai	1,003	0.56
kun	417	0.39	kun	342	0.73	sen	875	0.49
sen	395	0.37	nyt	220	0.47	kun	612	0.34
ole	337	0.32	sen	208	0.44	ole	584	0.33
vai	280	0.26	ole	144	0.30	hän	518	0.29
voi	212	0.20	voi	105	0.22	jos	415	0.23
nyt	192	0.18				voi	382	0.21
jos	175	0.16				eri	297	0.17
eri	104	0.10				eli	163	0.09
saa	103	0.10				saa	131	0.07
						nyt	126	0.07
						osa	110	0.06
11	3,417	3.22	7	2,205	4.70	13	6,399	3.60

4 graphemes

NEW	f	%	FIC	f	%	NON	f	%
että	898	0.84	että	355	0.76	että	1,838	1.03
ovat	484	0.45	niin	321	0.68	ovat	886	0.50
joka	453	0.42	minä	269	0.57	kuin	861	0.48
myös	450	0.42	kuin	267	0.57	myös	821	0.46
kuin	430	0.40	vain	181	0.39	sekä	653	0.37
niin	328	0.31	joka	147	0.31	joka	627	0.35
vain	255	0.24	mitä	140	0.30	vain	457	0.26
sekä	216	0.20	sitä	133	0.28	niin	455	0.26
sitä	211	0.20	sinä	113	0.24	tämä	367	0.21
tämä	194	0.18	ovat	102	0.22	sitä	359	0.20
mitä	167	0.16				olla	258	0.15
olla	146	0.14				koko	212	0.12
näin	143	0.13				mitä	211	0.12
eikä	123	0.12				eikä	200	0.11
koko	119	0.11				näin	191	0.11
aina	113	0.11				vaan	187	0.11
vaan	106	0.10				aina	179	0.10
						siis	178	0.10
						esim	140	0.08
						noin	130	0.07
						mikä	128	0.07
						taas	121	0.07
						jota	117	0.07
						nämä	115	0.06
						tuli	101	0.06
						tätä	101	0.06
17	4,836	4.55	10	2,028	4.32	26	9,893	5.57

5 graphemes

NEW	f	%	FIC	f	%	NON	f	%
mutta	523	0.49	mutta	368	0.78	mutta	639	0.36
ollut	241	0.23	hänen	150	0.32	ollut	399	0.22
hänen	236	0.22	ollut	129	0.27	jotka	390	0.22
jotka	226	0.21	olisi	125	0.27	olisi	318	0.18
olisi	205	0.19				hänen	306	0.17
sillä	200	0.19				siitä	305	0.17
siitä	185	0.17				jonka	299	0.17
vielä	180	0.17				tämän	270	0.15
tämän	171	0.16				koska	245	0.14
eivät	134	0.13				tässä	231	0.13
jonka	133	0.12				vielä	218	0.12
tässä	132	0.12				eivät	199	0.11
hyvin	126	0.12				kuten	197	0.11
jossa	124	0.12				hyvin	196	0.11
viime	122	0.11				tulee	190	0.11
juuri	120	0.11				sillä	188	0.11
siinä	117	0.11				siten	183	0.10
ennen	111	0.10				ennen	176	0.10
tulee	109	0.10				siinä	176	0.10
tällä	100	0.10				jossa	167	0.09
						usein	150	0.08
						ettei	142	0.08
						niitä	140	0.08
						joita	113	0.06
						antaa	110	0.06
						vasta	108	0.06
						tähän	107	0.06
						juuri	106	0.06
						pitää	105	0.06
						saada	103	0.06
						suuri	100	0.06
20	3,495	3.29	4	772	1.64	31	6,576	3.70

6 graphemes

NEW	f	%	FIC	f	%	NON	f	%
sitten	155	0.15	sitten	190	0.40	mukaan	445	0.25
paljon	151	0.14	kaikki	113	0.24	niiden	250	0.14
kaikki	141	0.13	olivat	101	0.22	olivat	210	0.12
olivat	120	0.11				siihen	189	0.11
suomen	118	0.11				kaikki	183	0.10
mukaan	116	0.11				aikana	179	0.10
vaikka	112	0.10				vuoden	176	0.10
vuoden	107	0.10				kanssa	171	0.10
kanssa	105	0.10				edellä	160	0.09
						varten	159	0.09
						sitten	150	0.08
						vaikka	145	0.08
						paljon	141	0.08
						vuoksi	136	0.08
						joiden	135	0.08
						olevan	126	0.07
						vuonna	126	0.07
						suomen	119	0.07
						varsin	116	0.07
						osalta	112	0.06
						avulla	109	0.06
						heidän	108	0.06
						samoin	106	0.06
						näiden	103	0.06
9	1,125	1.06	3	404	0.86	24	3,854	2.17

7 graphemes

NEW	f	%	FIC	f	%	NON	f	%
jälkeen	137	0.13				voidaan	436	0.25
voidaan	110	0.10				jälkeen	220	0.12
						lisäksi	198	0.11
						yleensä	172	0.10
						jolloin	148	0.08
						tällöin	146	0.08
						samalla	140	0.08
						silloin	140	0.08
						pykälän	124	0.07
						jumalan	122	0.07
						enemmän	119	0.07
						valtion	117	0.07
						päivänä	110	0.06
						saadaan	106	0.06
						tavalla	104	0.06
2	247	0.23	-	-	-	15	2,402	1.35

8 graphemes

NEW	f	%	FIC	f	%	NON	f	%
						huomioon	116	0.07
-	-	-	-	-	-	1	116	0.07

9 graphemes

NEW	f	%	FIC	f	%	NON	f	%
kuitenkin	195	0.18				kuitenkin	364	0.20
1	195	0.18	-	-	-	1	364	0.20

10 graphemes

NEW	f	%	FIC	f	%	NON	f	%
						yhteydessä	112	0.06
-	—	-	-	—	-	1	112	0.06

11 graphemes

NEW	f	%	FIC	f	%	NON	f	%
						perusteella	160	0.09
						tapauksessa	100	0.06
-	-	-	-	-	-	2	260	0.16
67	22,268	20.88	32	9,594	20.44	120	44,397	24.98

The samples of inflectional forms with at least 100 occurrences are percentually equally large in newspaper language and fictional language. The newspaper language contains 67 inflectional forms and a total of 22,268 graphological words, which accounts for 20.97 % of the graphological words in the whole category, whereas the sample of literary language has 32 inflectional forms, the total number of occurrences being 9,594, i.e. 20.44 % of the number of the graphological words. The sample of non-fictional language has a far higher percentage, the number of inflectional forms being 120 with 44,397 occurrences, i.e. 24.98 % of the total number of graphological words in this category. The NEW sample contains 7 inflectional forms with 2 graphemes, the FIC sample 8 and the NON sample only 6, but their frequency is high. They account for as much as 40.21 % of the graphological words in newspaper language, 43.62 % in fiction, but only 32.48 % in non-fiction. The proportion of graphological words with 3 graphemes is 15.34 % in newspaper language, 22.98 % in literary language and only 14.41 % in non-fiction, and the proportion with 4 graphemes is over one fifth in all the samples, i.e. 21.72 % in NEW, 21.14 % in FIC and 22.28 in NON. Graphological words with 5 graphemes account for 15.70 % in NEW, 14.81 % in NON but only 8.05 % in FIC, and the variation in the number of inflectional words involved is considerable, the figures being 20 - 4 - 31 in the above order. The proportion of graphological words with 6 graphemes is 5.05 % in NEW, 4.21 % in FIC and 8.68 % in NON. The language of fiction does not contain the required number of inflectional words longer than this for any further statistical analysis, but the proportion of graphological words with 7 graphemes

in the newspaper language is 1.11 % and that in the non-fiction sample 5.41 %. The 8-grapheme inflectional form *huomioon* enters the sample only in the case of non-fiction, with its 116 occurrences, 0.26 % of the graphological words. The 9-grapheme *kuitenkin* is percentually almost as common in newspaper language (195 occurrences = 0.88 %) as in the language of non-fiction (364 occurrences = 0.82 %). The occurrences of the 10-grapheme inflectional form *yhteydessä* are sufficiently numerous in the language of non-fiction (112 = 0.25 %), as are those of two inflectional forms with 11 graphemes, *perusteella* (160 occurrences) and *tapauksessa* (100 occurrences, accounting together for 0.59 % of the sample).

The high frequencies of the short inflectional forms in a sample lead to a situation in which the number of inflectional forms of medium length included in a sample based on a minimum of 100 occurrences is extremely small. As mentioned above, the mean number of graphemes per word in the total set of graphological words is 7.58 in newspaper language, 6.39 in fiction and 7.8 in non-fiction, which at the grapheme level leads to the fact that a sample containing a little over 20 % of the total graphological words in the category for only some 10 % of the graphemes. The graphological words of newspaper language in the sample thus contain only 75,210 graphemes (9.34 % of the 804,842), those in the fiction sample 29,381 graphemes (9.79 % of the 300,093) and those in the non-fiction sample 168,541 graphemes (12.16 % of the 1,386,273). The mean number of graphemes per graphological word is 3.38 in NEW, 3.06 in FIC, and 3.80 in NON. The mutual orders and percentages of graphemes in all three samples of graphological words with over 100 occurrences differ in a number of ways from the corresponding systems of graphemes in the categories as a whole, as shown in the table below, which also contains the percentages of graphemes in the basic categories to facilitate comparison (NEW %, FIC %, and NEW %).

Table 14 clearly indicates that the set of graphemes in the graphological words with the highest frequencies differs from the mean in many ways and to a great extent. It could be assumed that such short, frequent graphological words, which are sometimes highly characteristic of certain styles (e.g. *ja*, *siis*, *myös*, etc.), would have a crucial influence on the mutual order of graphemes in the various stylistic classes, but the above calculations do not support such an assumption. Take the grapheme a, for instance, the occurrence of which is highly significantly greater than average in the total sample of newspaper language (+0.48 percentage point), but almost a percentage point lower than this in the sample with frequencies of over 100, even though the graphological word *ja*, for instance, is very common.

TABLE 14. Graphemes of inflectional forms with a minimum of 100 occurrences in the language of newspapers, fiction and non-fiction, in order of frequency.

	NEW sample (NEW %)				FIC sample (FIC %)				NON sample (NON %)			
		f	%			f	%			f	%	
1.	n	9,220	12.26	(8.72)	n	4,229	14.39	(9.09)	a	20,387	12.10	(11.83)
2.	a	8,511	11.32	(12.10)	i	3,830	13.04	(11.30)	n	20,138	11.95	(8.55)
3.	o	7,682	10.21	(5.34)	a	2,886	9.82	(11.74)	i	17,361	10.30	(10.50)
4.	i	7,504	9.98	(10.64)	t	2,651	9.02	(9.34)	o	15,263	9.06	(5.28)
5.	t	6,013	7.99	(9.59)	o	2,453	8.35	(4.77)	t	13,662	8.11	(9.96)
6.	j	5,767	7.67	(2.03)	e	2,452	8.35	(7.79)	e	13,405	7.95	(8.31)
7.	e	5,707	7.59	(8.19)	j	2,170	7.39	(1.89)	s	10,421	6.18	(8.16)
8.	ä	4,576	6.08	(4.40)	ä	1,652	5.62	(5.58)	ä	10,138	6.02	(4.48)
9.	s	3,807	5.06	(7.66)	l	1,322	4.50	(5.85)	j	9,791	5.81	(1.84)
10.	k	3,631	4.83	(5.33)	s	1,257	4.28	(6.96)	l	8,203	4.87	(5.56)
11.	l	3,380	4.49	(5.82)	u	1,106	3.76	(5.18)	k	8,021	4.76	(5.26)
12.	u	2,496	3.32	(4.99)	k	1,095	3.73	(5.75)	u	5,534	3.28	(5.08)
13.	v	2,348	3.12	(2.47)	h	792	2.70	(2.31)	v	4,703	2.79	(2.58)
14.	m	1,861	2.47	(3.38)	m	777	2.64	(3.18)	m	3,977	2.36	(3.28)
15.	h	897	1.19	(1.90)	v	489	1.66	(2.40)	d	1,689	1.00	(0.93)
16.	y	768	1.02	(2.47)	y	220	0.75	(1.58)	h	1,652	0.98	(1.68)
17.	ö	450	0.60	(0.46)	-				y	1,551	0.92	(1.88)
18	r	224	0.30	(2.36)	-				ö	967	0.57	(0.53)
19.	d	217	0.29	(0.80)	-				r	938	0.56	(2.34)
20.	p	151	0.20	(1.71)	-				p	740	0.44	(1.65)
		75,210	100.0			29,381	100.0			168,541	100.0	

Correspondingly, the frequency of a in the restricted sample of literary language is almost two percentage points lower than in the total sample, whereas, surprisingly, it is 0.27 percentage points higher in the NON sample. A second suitable example is the grapheme ä, which occurs very frequently in the language of fiction (deviation from the mean +0.77 percentage points), but is relatively rare in the language of newspapers and non-fiction (deviations from the mean -0.41 and -0.33 percentage points, differences statistically highly significant). The ratios in the sets of graphemes with over 100 occurrences are nevertheless almost

the opposite, the graphological words with high frequencies in fiction have almost an equal number of ä:s as the total set of words, while the corresponding number in the samples of newspaper language and non-fiction is larger by almost 1.5 percentage points. The frequencies of n and j are conside-rably higher in the restricted samples than in the total samples, which must be due to the graphological words *on*, *ne*, *hän*, *kun*, *sen*, *ja*, and *jo*. It is thus obvious that variations in the inflectional forms with the highest occurrences are not responsible for the overall variations in the frequencies of graphemes in the various stylistic categories.

In his investigation "Suomen kielen dynamiikkaa" ("Dynamics of the Finnish language"; 1972), Vilho Setälä examines the frequency relations of sounds on the basis of theories developed by G.K. Zipf (Zipf 1965), according to which the frequency of sounds is universally connected with ease of pronunciation, in that the easiest to pronounce are the most common. Without any empirical investigations, Setälä drew the conclusion that the front vowels of Finnish, ä, y, and ö, require more intensive articulation than the back vowels a, u, and o, and that this is also reflected in the frequency of the corresponding graphemes. He also assumed that the articulation of m requires more energy than that of n, and that r is rare in Finnish since it requires a lot of energy, while in German, for instance, it is much more frequent, being weaker (ibid. 37 - 39, 42). Setälä states that other phonemes are more difficult to classify according to articulation, so that he does not comment on whether the plosive t, which occupies the 3rd position in the overall table (n^1), requires less energy than k, which occupies 9th position, or p in 18th position.

"Zipf's Law", according to which the number of different words in any long stretch of text in any language is constant - and its corollary, that the frequency of graphemes should also be constant - does not seem to hold good in the present material. Samples of this size are apparently not sufficiently large for the law to hold good. George A. Miller, editor of the latest edition of Zipf's linguistically most significant work, states in his foreword that according to mathematical laws, Zipf's famous vocabulary curves hold good in the language produced by monkeys using a typewriter, if the sample is extensive enough.

9. Bound morphemes and graphemes

The extensive material used here and that used by Vilho Setälä seem to indicate that morphological issues crucially affect the frequency of graphemes in the

Finnish language. Take the statistics compiled by Setälä and Saukkonen, for instance: n is the last grapheme in graphological words in the New Testament in 28.08 % of cases in Setälä's material, and a or ä in 36.74 % (1972: 23), while n is the last grapheme in 33 % of the graphological words in Saukkonen's rough statistics, a in 27 % and ä in a further 10 % (1977: IX).

The frequencies of graphemes were also calculated here for the languages of newspapers, fiction, and non-fiction starting from the end of the graphological word and proceeding backwards one position at a time. The seven most frequent graphemes are often the most common at the ends of graphological words. The frequencies of the graphemes a + ä = A, i, t, n, e, s and l in the last four positions in graphological words are presented in the table below. The graphemes are tabulated by main textual categories, and the figures are expressed as percentages of the occurrences of the same grapheme in the category as a whole. The column designated (-1) applies to the last grapheme in the graphological word, column (-2) the second last, (-3) the third last and (-4) the fourth last.

Table 15 above indicates that approximately half of all occurrences of n in each main category are at the end of a graphological word, 30 % of those of A and approx. 12 % and 11 % of the cases of i in the categories NEW and NON respectively, but almost twice as many in FIC, i.e. over 21 %. Correspondingly, 23.33 % - 24.77 % of the total occurrences of e are located in the second last position in graphological words, 25 % of s in the third last position, and approx. 20 % of i in the fourth last position.

The frequencies of the seven most common graphemes in the last four positions are presented in table 16 as percentages of the graphological words in the main categories, i.e. the figures indicate the percentages of words in these categories that have the given graphemes in these positions. The columns are arranged in the same manner as in the previous table.

TABLE 15. Graphemes occurring in the last four positions of grapho-logical words in the language of newspapers, fiction, and non-fiction in order of frequency relative to the total number of occurrences of the same grapheme in the textual category.

	(-1)	%	(-2)	%	(-3)	%	(-4)	%
NEW								
1.	n	50.98	e	23.33	s	25.54	i	20.12
2.	A	29.48	t	17.80	l	19.52	e	16.37
3.	i	12.18	l	16.39	t	13.77	A	10.51
4.	t	9.46	s	14.89	n	11.04	t	9.84
5.	e	8.56	A	13.30	i	8.86	s	8.09
6.	s	4.60	i	12.80	e	7.82	l	7.98
7.	l	0.22	n	5.03	A	7.23	n	6.70
FIC								
1.	n	49.50	e	24.77	s	24.37	i	20.84
2.	A	31.36	l	21.76	l	23.17	e	17.83
3.	i	21.72	t	19.82	t	16.33	A	11.55
4.	t	14.25	s	19.24	n	14.56	t	11.06
5.	e	12.66	A	15.82	i	11.25	l	10.63
6.	s	5.78	i	14.40	e	9.55	s	9.63
7.	l	0.28	n	7.91	A	8.44	n	7.23
NON								
1.	n	52.82	e	24.27	s	25.55	i	19.60
2.	A	29.09	t	16.99	l	18.11	e	16.42
3.	i	11.03	l	15.45	t	13.72	t	9.87
4.	t	8.34	s	13.74	n	9.26	A	9.10
5.	e	6.31	A	12.67	i	8.44	s	8.68
6.	s	4.61	i	12.43	e	8.03	l	7.35
7.	l	0.29	n	4.59	A	7.23	n	5.88

TABLE 16. Graphemes occurring in the last four positions of grapho-logical words in the languages of newspapers, fiction, and non-fiction in order of frequency relative to the total number of graphological words in the textual category.

	(-1)	%	(-2)	%	(-3)	%	(-4)	%
NEW								
1.	A	36.70	A	16.55	s	14.77	i	16.15
2.	n	33.53	e	14.42	t	9.96	A	13.08
3.	i	9.77	t	12.88	A	9.00	e	10.11
4.	t	6.84	i	10.27	l	8.57	t	7.12
5.	e	5.29	s	8.61	n	7.26	s	4.68
6.	s	2.66	l	7.20	i	7.11	n	4.40
7.	l	0.09	n	3.31	e	4.83	l	3.51
FIC								
1.	A	34.71	A	17.51	s	10.85	i	15.05
2.	n	28.77	e	12.33	t	9.75	A	12.78
3.	i	15.69	t	11.83	A	9.34	e	8.88
4.	t	8.51	i	10.40	l	8.67	t	6.60
5.	e	6.30	s	8.56	n	8.46	s	4.29
6.	s	2.57	l	8.15	i	8.12	n	4.20
7.	l	0.10	n	4.60	e	4.75	l	3.98
NON								
1.	A	37.00	A	16.11	s	16.26	i	16.05
2.	n	35.22	e	15.73	t	10.66	A	11.58
3.	i	9.03	t	13.20	A	9.20	e	10.64
4.	t	6.48	i	10.18	l	7.86	t	7.67
5.	e	4.09	s	8.74	i	6.91	s	5.52
6.	s	2.93	l	6.70	n	6.17	n	3.92
7.	l	0.13	n	3.06	e	5.21	l	3.19

The frequencies of graphemes at the ends of graphological words are compared in the above table with the total number of graphological words in each category. Inflectional forms terminating in A are by far the most common in all the categories (34.71 - 37.00 %), while the number of words terminating in n is almost as large (28.77 - 35.22 %). The proportion of those ending in i is already below 10 % in the categories NEW and NON, however, although it is as high as 15.69 % in FIC. Graphological words most commonly end in the sequence i s A A.

It is important to bear in mind when interpreting Tables 15 and 16 that while position (-1) really is always the last one in a graphological word, positions (-2), (-3) and (-4) can sometimes be initial graphemes of words of that length, even though the mean number of graphemes in graphological words in the textual categories described here is approximately 7 (NEW 7.58, FIC 6.39, and NON 7.80). The percentages presented in Table 15 are accurate values, since they were calculated from the total numbers of occurrences of the corresponding grapheme in each category of written language, but those in Table 16 are only approximate, being calculated from the number of occurrences of graphological words without taking into account their length, i.e. the error caused by short graphological words. Even so, they do provide information on the structure of the sequences of graphemes in a terminal position in graphological words and the frequencies of graphemes occurring in suffixes. The graphs below were drawn up to illustrate this by showing the levels of occurrence of the seven most common graphemes (A = a + ä) in the last four positions in graphological words on the basis of the percentages presented in Table 16.

The arrangement of columns with respect to position in the graphological word is the same as above. The main categories NEW, FIC, and NON can be compared by means of adjacent columns.

FIG. 1. A in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

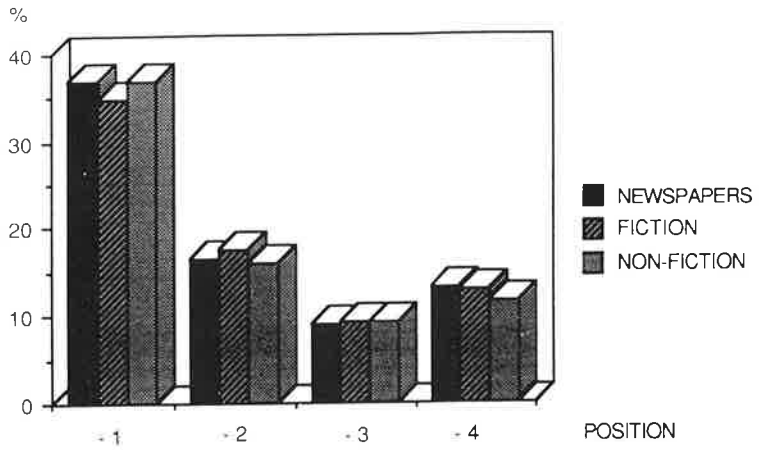


FIG. 2. n in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

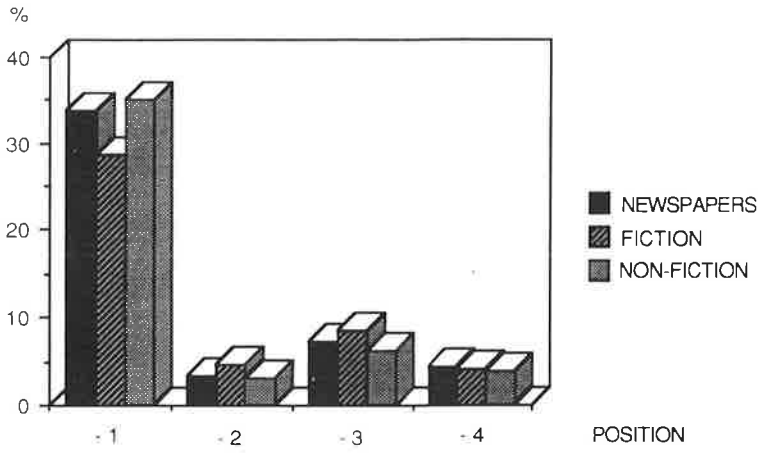


FIG. 3. i in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

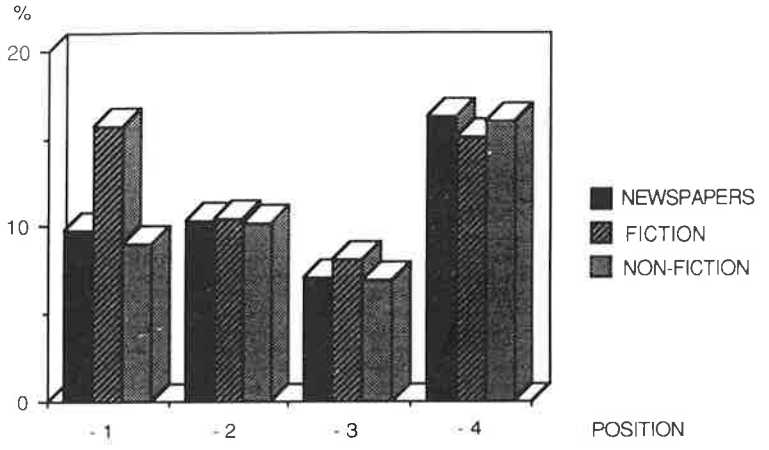


FIG. 4. t in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

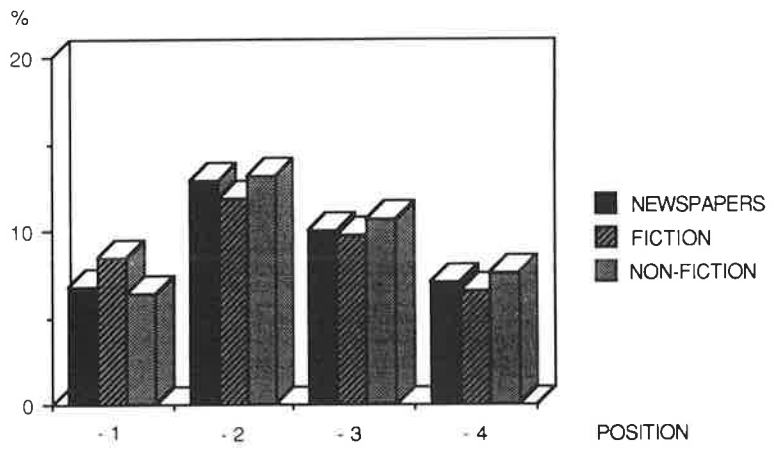


FIG. 5. e in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

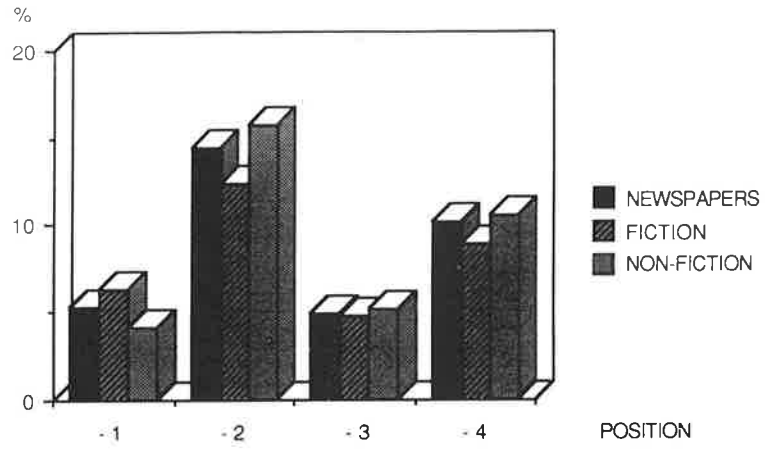


FIG. 6. s in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.

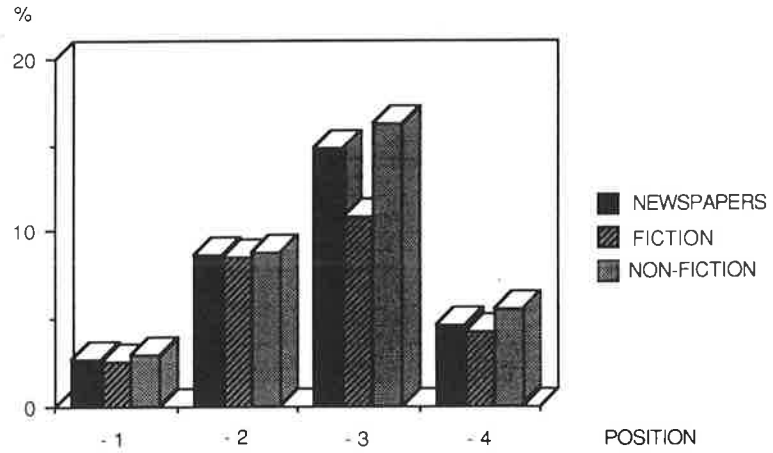
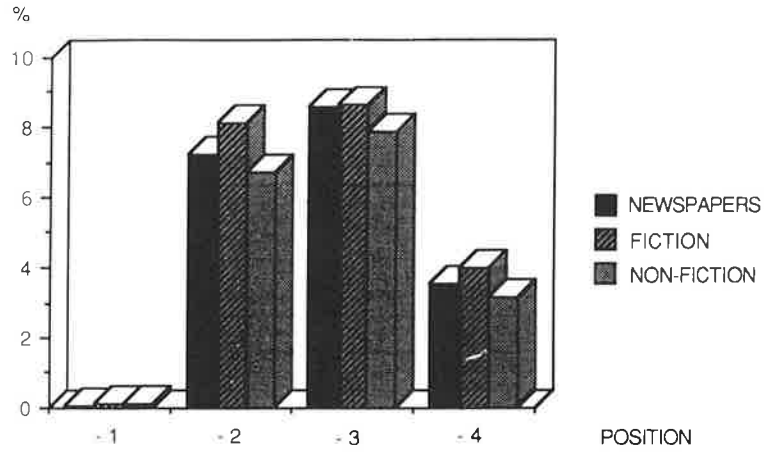


FIG. 7. l in the last four positions in the graphological word in the language of newspapers, fiction, and non-fiction.



The seven most frequently occurring graphemes in Finnish are a, i, t, n, e, s, and l, and it is these (+ä) that are also the most frequent in the last four positions in graphological words. The most common stems in the basic words of our nominal vocabulary terminate in i or A, and derivatives are also fairly commonly formed by means of the suffixes s, nen, jA, lA and nA (Karlsson 206 - 207). These most frequent graphemes are crucial elements in bound, overt morphemes. Thus the standard case endings are most often composed of open vowels and dental consonants, and often also contain e and i (A, llA, lTa, ssA, stA, tA, ttA, nA, lle, n, den, ten, tten, en, ksi, kse, Vn, hVn, seen, siin). The morphs indicating person and possession (n, t, mme, tte, mme, nne, nsA) and those indicating tense, mood, and number (i, isi, ne, kO, kAA, i, t) are again mainly composed of these most frequent elements (Karlsson 306 - 307). The most prominent variations in the frequencies of graphemes thus apparently reflect variations in the textual frequencies of such bound, overt morphs in various contexts. It also seems that the grapheme sequences in terminal positions in graphological words also determine the mutual order of frequency of these most common graphemes at least (and thus also that of the phonemes of Finnish). The dental plosive t, which occupies the 3rd position in the table of graphemes, is thus not universally any easier to produce than the palatal plosive k in the 9th position or the labial plosive p in the 18th position, but it is by far the most frequent of the three in inflectional elements, endings and suffixes, and p correspondingly the rarest.

Similarly, Vilho Setälä's assumption that the rarity of m as compared with n or the low frequency of the Finnish r as compared with the German r is due to the large expenditure of energy on articulation is thus unnecessary speculation. Their numerical inferiority in Finnish can be explained simply by the fact that they are rare in word final suffixes and inflections, although the reasons for this do not belong to the scope of the present investigation.

When the detailed analysis and statistical tabulation of grammatical categories in all the textual classes included in the Oulu Corpus is complete, it will also be possible to examine in greater detail the reasons for the variations in the frequencies of graphemes in both large and small text classes and in various contexts.

APPENDICES

The appendices to this monograph include tables of graphemes in the remaining main textual categories (RAD = radio language, SPK = free standard spoken language and PARL = parliamentary language).

Appendix 1. Radio Language (style categories 70 - 81)

The sub-categories represent standard spoken language, which is nevertheless a fairly diverse class. The sample contains 78,964 graphological words and 567,944 graphemes, the former of which are shorter than the average, the mean number of graphemes per word being 7.19. E^1 = deviation of the frequencies of the graphemes in radio language from the mean (n^1) in percentage points.

TABLE 17. Graphemes in radio language (style categories 70 - 81).

	(n^1)	RAD	f	%	E^1
1.	(a)	a	65,156	11.47	-0.15
2.	(i)	i	60,569	10.66	-0.05
3.	(t)	t	55,423	9.76	-0.12
4.	(n)	n	49,510	8.72	+0.05
5.	(e)	e	47,360	8.34	+0.13
6.	(s)	s	44,205	7.78	-0.08
7.	(l)	l	33,463	5.89	+0.13
8.	(o)	o	30,520	5.37	+0.06
9.	(k)	k	30,072	5.29	+0.02
10.	(u)	ä	28,739	5.06	+0.25
11.	(ä)	u	929	4.92	-0.08
12.	(m)	m	27,536	3.79	+0.28
13.	(v)	v	21,135	2.31	-0.14
14.	(r)	r	13,563	2.04	-0.12
15.	(j)	j	11,747	1.89	-0.04
16.	(h)	h	10,320	1.82	0.00
17.	(y)	y	10,064	1.77	-0.04
18.	(p)	p	10,627	1.70	+0.04
19.	(d)	d	9,658	0.82	-0.02
20.	(ö)	ö	4,221	0.39	-0.08
21.	(g)	g	2,570	0.10	-0.01
22.	(b)	b	284	0.05	0.00
23.	(f)	f	234	0.04	-0.01
24.	(c)	c	27	0.00	-0.03
25.	(w)	w	15	0.00	-0.01
26.	(å)	å	8	0.00	0.00
27.	(q)	q	-	-	-
			567,944	100.00	

Radio language does not seem to deviate from the mean to any significant extent. The order of frequency has only one deviation: ä is in 10th position and is markedly more common than u. The frequencies of e, l, ä, and m are well above the average, while a, t, v, and r are below average.

Appendix 2. Free standard spoken language (style category 90)

Free standard speech makes up the smallest of the main categories, and is included here mainly as reference material. The number of graphological words in this category is 12,031 and that of graphemes 71,230, the mean number of graphemes per word being only 5.92, i.e. 1.44 less than on average. E^1 = deviation of the frequencies of graphemes in free standard spoken language from the mean (n^1) in percentage points.

TABLE 18. Graphemes in free standard speech (category 90).

	(n^1)	FSS	f	%	E^1
1.	(a)	i	8,617	12.10	+1.39
2.	(i)	a	7,717	10.83	-0.79
3.	(t)	t	7,070	9.93	+0.05
4.	(n)	n	6,361	8.93	+0.26
5.	(e)	e	5,643	7.92	-0.29
6.	(s)	s	5,468	7.68	-0.18
7.	(l)	l	4,276	6.00	+0.24
8.	(o)	o	4,028	5.65	+0.34
9.	(k)	ä	3,757	5.27	+0.46
10.	(u)	k	3,624	5.09	-0.18
11.	(ä)	u	3,170	4.45	-0.55
12.	(m)	m	2,956	4.15	+0.64
13.	(v)	j	1,682	2.36	+0.43
14.	(r)	v	1,528	2.15	-0.30
15.	(j)	h	1,297	1.82	0.00
16.	(h)	y	1,224	1.72	-0.09
17.	(y)	r	1,114	1.56	-0.60
18.	(p)	p	998	1.40	-0.26
19.	(d)	d	319	0.45	-0.39
20.	(ö)	ö	264	0.37	-0.10
21.	(g)	g	64	0.09	-0.02
22.	(b)	f	27	0.04	-0.01
23.	(f)	b	26	0.04	-0.01
24.	(c)	-	-	-	-
25.	(w)	-	-	-	-
26.	(å)	-	-	-	-
27.	(q)	-	-	-	-
			71,230	100.00	

The mutual order of graphemes in this category deviates from the mean at many points. The most conspicuous difference is found at the top of the table, where the most common grapheme is not a but i, the frequency of which is as much as 1.39 % above the average. The high frequency of i seems to be characteristic of spoken language, since it is very much more common than a both in dialects in general (Pajunen - Palomäki 1984: 73 - 89) and in the radio interviews and discussions in the present material (categories 80 and 81); i also occupies the leading position in the language of drama for young people within the fiction category (category 31), and, oddly enough, in a subcategory of non-fiction, namely "Reviews of non-fiction" (category 06). The high occurrence of i could be due to the frequent use of active indicative verb forms and the past tense of passive finite forms in these categories. Of the other graphemes, ä is two positions higher than on average, while k and u are located one position lower than normal. o and m occupy their usual positions, although their occurrence in spoken language is considerably above the average. Great variation also occurs in the next five positions, from 13th to 17th. r does not seem to occur very frequently, being three positions lower than usual, while j is more common, being two positions higher than usual, and the occurrence of d is considerably below average, as can be expected.

Appendix 3. Parliamentary language (categories 91 and 92)

Category 91, part of the material of parliamentary language compiled by Esko Vierikko, contains free speech from interviews with representatives, which has been transcribed roughly but nevertheless includes sandhi assimilations and often post-aspirations as well. Category 92 contains official parliamentary language, mainly from plenary sessions, 2/3 of it being transcribed from tapes and 1/3 drawn from the official parliamentary records and normalized by the secretaries. The total number of graphological words is 112,606 and that of graphemes 804,771, the mean number of graphemes per word being 7.15. E^1 = deviation of the frequencies of the graphemes in parliamentary language from the mean (n^1) in percentage points.

TABLE 19. Graphemes in parliamentary language (categories 91 - 92).

	(n^1)	PARL	f	%	E^1
1.	(a)	a	87,939	10.93	-0.69
2.	(i)	i	87,173	10.83	+0.12
3.	(t)	t	82,960	10.31	+0.43
4.	(n)	n	69,399	8.62	-0.05
5.	(e)	e	65,639	8.16	-0.05
6.	(s)	s	64,002	7.95	+0.09
7.	(l)	l	47,335	5.88	+0.12
8.	(o)	o	43,830	5.45	+0.14
9.	(k)	ä	42,384	5.27	+0.46
10.	(u)	k	40,834	5.07	-0.20
11.	(ä)	u	39,415	4.90	-0.10
12.	(m)	m	31,208	3.88	+0.37
13.	(v)	v	18,873	2.35	-0.10
14.	(r)	j	15,941	1.98	+0.05
15.	(j)	y	15,360	1.91	+0.10
16.	(h)	r	14,617	1.82	-0.34
17.	(y)	h	14,613	1.82	0.00
18.	(p)	p	11,451	1.42	-0.24
19.	(d)	d	7,101	0.88	+0.04
20.	(ö)	ö	3,982	0.49	+0.02
21.	(g)	g	371	0.05	-0.06
22.	(b)	b	165	0.02	-0.03
23.	(f)	f	134	0.02	-0.03
24.	(c)	c	23	0.00	-0.03
25.	(w)	å	14	0.00	0.00
26.	(å)	w	7	0.00	-0.01
27.	(q)	q	1	0.00	0.00
			804,771	100.00	

What most attracts attention in the list of graphemes in parliamentary language is the low frequency of a, which is 0.69 percentage points below average, while the occurrences of t, ä, and m are considerably above average. The high frequency of ä is also conspicuous both in the language of the New Testament and in the present categories of literary and radio language and free standard speech. One explanation for this could be the frequent use of personal pronouns, for the occurrence of h is also statistically highly significantly above average in the language of the New Testament and the present fiction category, although it is exactly at the mean value in radio language, free standard speech and parliamentary language. The frequency of r is also well below the mean in parliamentary language, being two positions lower than the normal.

LITERATURE

ALLÉN, Sture (1970): *Nusvensk frekvensordbok.* Stockholm.

HÄKKINEN, Kaisa (1978): *Eräistä suomen kielen äännerakenteen luonteenomaisista piirteistä.* Käsikirjoitteena oleva lisensiaatintyö. Säilyt-teillä Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksessa.

JÄRVIKOSKI, Olli (1985): *Suomen kielen foneemien ja grafeemien frekvensseistä.* Virittäjä 89. Helsinki.

KARLSSON, Fred (1983): *Suomen kielen äänne- ja muotorakenne.* Juva.

NIEMIKORPI, Antero (1972): *Saneiden pituus.* Käsikirjoitteena oleva lisensiaatintyö. Säilytteillä Oulun yliopiston suomen ja saamen kielen laitoksessa.

PAJUNEN, Anneli & PALOMÄKI, Ulla (1984): *Tilastotietoja suomen kielen rakenteesta 1.* Kotimaisten kielten tutkimuskeskuksen julkaisuja 30. Valtion painatuskeskus. Helsinki.

PESONEN, Jaakko (1971): *Sanamuodot ja niiden kirjainrakenne suomenkielisessä sanomalehtitekstissä.* Research reports no. 6/1971. Department of special education. University of Jyväskylä.

PÄÄKKÖNEN, Matti (1973): *Tilastotietoja suomen yleiskielen grafeemeista. Suomalais-ugrilaisen Seuran Aikakauskirja 72.* Helsinki.

SAUKKONEN, Pauli, HAIPUS, Marjatta, NIEMIKORPI, Antero & SULKALA, Helena (1979): *Suomen kielen taajuussanasto.* Porvoo.

SAUKKONEN, Pauli (1977): *Nykysuomen saneiston yleisyystilastoa saneenloppuisessa aakkosjärjestyksessä.* Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 9. Oulu.

SAUKKONEN, Pauli (1982): *Oulun korpus.* 1960-luvun suomen yleiskielen tutkimusmateriaali. Mikrokortit toimittaneet Marjatta Haipus ja Tapani Sulkala. Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 1. Oulu.

SETÄLÄ, Vilho (1972): *Suomen kielen dynamiikkaa.* Suomi 116:3. Helsinki.

ZIPF, George K. (1965): *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Cambridge, Massachusetts: Massachusetts Institute of Technology .

Glottometrika 14, 1993, 54-68

Phoneme Counts

Marginal remarks to Pääkkönen's article

Gabriel Altmann, Bochum

Counting of elementary linguistic or graphemic units is a fruitful tradition of Finnish linguistics. Unfortunately, the results were always published in Finnish and remain unavailable in spite of the fact that everybody can read numbers. M. Pääkkönen redresses this evil in that he presents in this volume an extract from his extensive book (Grafeemit ja konteksti. Helsinki 1990) and compares his own counts with those of other Finnish scholars.

Data of this kind are of inestimable value also to theoretical quantitative linguistics since they allow us to consider some general problems. The following points are not meant as criticism since Pääkkönen pursued another aim in his work, namely a possible text characterization by means of letter frequencies, i.e. he searched for differences while our goals are quite the opposite: we search for constancies and our view is purely methodological. We merely use his data for discussing some problems.

Size

The enormous size of these data - almost four million of letters - brings both advantages as well as disadvantages.

The advantages - if they turn to be so - consist in the possibility of

(1) computing global language characteristics,

(2) a more confident comparison of Finnish with other languages, and

(3) testing some quantitative models.

However if the frequencies are not stable even with an enormous sample size then all the advantages mentioned above are illusory; nevertheless, the data can be used for characterization of individual texts.

The disadvantages are as follows:

(1) The overall data are mixtures of many different samples. Perhaps mixing renders some language properties more stable but there are problems with fitting models e.g. rank-frequency distributions to such data.

(2) If one tests the differences between the proportions of individual letters in two samples by means of a *t*-test, one obtains as a rule highly significant results in spite of the fact that the differences are actually extremely small. This is caused by the enormous size of the samples; it is well known that any difference of this kind can be made significant if one uses sufficiently large samples. This gives the impression that either the frequencies do not achieve stability even if a sample consists of millions of items, or that the correctly used statistical method is problematic. Though large sample sizes are no disadvantage, quite the contrary, it is statistics that brings the headache.

Stability

As was shown by Orlov (1982) there are no populations in language. In quantitative linguistics language is not considered as a homogeneous entity with firm boundaries and 'natural constants'. The only things that are constant in it are the relations between its variable properties, which are arranged in control cycles (cf. Köhler 1986, Altmann 1987). Probably any complete text is a population of its own; there are not even constant text properties for 'all works of an author'. Consequently there can be no entity called e.g. 'the language of newspapers' displaying constants for all possible properties. Nevertheless there can be classes of texts held together by one or more features. Thus one hopes that a large mixture of many samples drawn from the same class brings stability for at least some of the properties under consideration. However a problematic statistical test (cf. Table 4 in Pääkkönen) can upset this assumption. Consequently we are committing an error somewhere. Either we use an inadequate statistical method or we must search for other kinds of stability or both. If not even millions of occurrences of a letter stabilize its relative frequency, what is stable with letter frequencies? There are several possibilities.

(a) The first possibility is that even though the frequencies of equal letters in two large samples are significantly different according to the *t*-test, it is the problem of the test and not of reality.

(b) The second possibility is that even if a different test would correctly

show significant differences in proportions, merely *the distribution of letters as a whole* is homogeneous in the samples compared.

(c) The third possibility is that merely the *ranks* of the letters stabilize gradually; however it need not be so even with two and a half million letters (cf. Pääkkönen, Table 1).

(d) The fourth possibility is that merely a *global measure for the whole distribution* is stable.

(e) We can even imagine that it is the *relation between the individual frequencies* that is stable, i.e. the frequency of the most frequent letter is in the same (numerical) relation to the frequency of the second most frequent letter, as the second to that of the third most frequent one, etc. This must hold true for all long texts (large samples) independently of the rank of the individual letters.

Let us consider possibilities (b) to (e) one by one, indicating these letters by numbers (I) to (IV).

Homogeneity

(I) In order to test the homogeneity of two multinomial samples one usually uses the chi-square (or some equivalent) test according to the formula

$$(1) \qquad X^2_{K-1} = \frac{n^2}{n_{.1}n_{.2}}\left[\sum_{i=1}^{K}\frac{n_{i1}^2}{n_{i.}} - \frac{n_{.1}^2}{n}\right]$$

where n_{i1} = frequency in cell i of the first column
$n_{.1} = \Sigma n_{i1}$ (i = 1,2,...,K) (the sum of the first column)
$n_{.2} = n - n_{.1}$ (the sum of the second column)
$n = n_{.1} + n_{.2}$ (total sum)
K = number of rows (here letters).

After reordering the letters in Pääkkönen's Table 1 for 'transcribed spoken language' (S1) and 'written language' (S2) we obtain Table 1 below.

The result according to (1) is $X^2 = 4832.2849$ with 26 degrees of freedom,

Table 1. Testing homogeneity of data S1 and S2 of Pääkkönen

	S1	S2
a	160812	296538
i	156359	265007
t	145442	243269
n	125270	215911
e	118642	204445
s	113675	195675
l	85074	141553
o	78378	130545
ä	74880	114254
k	74530	132990
u	70514	126164
m	55700	82272
v	33536	62780
j	28370	47591
r	27294	57822
y	26648	44668
h	26230	45503
p	22076	43282
d	12078	21070
ö	6467	12188
g	1005	3146
b	475	1593
f	395	1539
c	52	1041
w	22	30
å	20	30
q	1	25

which is considered as an indication of heterogeneity. Now since the chi-square grows linearly with increasing sample size, with large samples one rather uses the contingency coefficient $C = \sqrt{X^2/N}$, yielding here C = 0.035 showing that the samples are homogeneous.

So the situation when basing the conclusions on large samples is not unambiguous.

(II) We can of course dispense with the frequencies and consider merely the rank order of letters in the two samples. As can be seen, the rank order of letters according to their frequency is not identical in the samples under consideration. Taking the order in S1 as the basic one we obtain the ranks as given in Table 2. Our problem is whether the given disorder can be considered merely as random fluctuation of letter ranks or as something caused by unknown factors.

There are a number of procedures for evaluating the agreement of these two rankings. The simplest one is to relate the actual differences of ranks to their maximum. In order to measure the stability of ranks we may compute Spearman's rank correlation coefficient

$$(2) \qquad r = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N}$$

where d_i is the difference between the ranks of the i-th letter. For our example we obtain

$$r = 1 - \frac{6(14)}{27^3 - 27} = 0.9957.$$

Or, we can use Kendall's rank correlation coefficient

$$(3) \qquad \tau = \frac{2S}{N(N-1)}$$

where S is the sum of +1 and -1 scores for correct and incorrect rankings in the second sample (S2). It is easy to compute by hand. For Table 2 we obtain

$$\tau = \frac{2(341)}{27(26)} = 0.9715.$$

In both cases we obtain a strong agreement between the ranks in the two samples under consideration.

Table 2. Rank order of letters according to their frequency

	a i t n e s l o ä k u m v j r y h p d ö g b f c w å q
S1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
S2	1 2 3 4 5 6 7 9 11 8 10 12 13 15 14 17 16 18 19 20 21 22 23 24 25 26 27
d_i	1 2 -2 -1 1 -1 1 -1

We also can test the stability of ranks using all individual samples (S1 and S2 are merely their sums). In his book, Pääkkönen (1990: 85) published a table of letter rankings for 60 text collections which is given in Table 3. Ties were taken into account but they are present merely in the last columns if the letters concerned did not occur in the sample at all. To measure the agreement of more than two orders, we can use Kendall's coefficient of concordance (cf. Siegel 1956)

$$(4) \qquad W = \frac{s}{k^2(N^3 - N)/12 - k\sum T}$$

where N = number of letters (here 27)
k = number of compared sets (here 60)
$s = \Sigma(r_j - \Sigma r_j/N)^2$
r_j = sum of all ranks of the letter j
$T = \Sigma(t^3 - t)/12$, t being the number of tied ranks.

For Table 3 we obtain W = 0.9890. A transformation to chi-square yields X^2 = k(N-1)W = 60(26)0.9890 = 1542.77 indicating that there is a very high agreement of ranks in these 60 samples.

Table 3. Rankings of Finnish letters from 60 samples by Pääkkönen

a i	t	n	e	s	l	o	k	u	ä	m	v	r	h	j	y	p	d	ö	g	b	f	c	w	å	q
1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	15	17	18	19	20	21	22	23	24	25	26.5	26.5
1	2	3	4	5	6	7	8	10	9	11	12	14	13	15	16	17	18	19	20	21	24	22	23	25	26	27
1	2	3	4	5	6	7	11	8	9	10	12	13	15	16	14	18	17	19	20	21	23	22	24	25	26.5	26.5
1	3	2	5	4	6	7	11	10	9	8	12	13	16	15	14	17	18	19	20	21	23	22	26	26	24	26
1	2	3	4	5	6	9	8	7	10	11	12	13	14	15	16	18	17	19	20	21	22	24	23	25	27	26
2	1	3	4	6	5	9	7	8	10	11	12	13	14	16	17	15	18	19	20	21	23	22	24	25	27	26
1	2	3	4	5	6	7	9	8	10	11	12	14	13	15	16	17	18	19	20	21	22	24	23	25	27	26
1	2	4	3	5	6	7	10	8	9	11	12	13	14	15	16	18	17	19	20	21	22	23	24	25	26	27
1	2	3	5	4	6	7	8	9	10	11	12	13	14	15	17	16	18	19	20	21	22	24	23	26	26	26
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26.5	26.5
1	2	3	4	5	6	7	10	8	9	11	12	14	13	15	16	18	17	19	20	21	22	23	24	25	26	27
1	2	3	4	5	6	7	9	8	10	11	12	13	15	16	14	17	18	19	20	21	22	23	25	24	26.5	26.5
1	2	3	4	5	6	7	8	9	10	11	12	14	13	15	16	17	18	19	20	21	22	23	24	25	26	27
1	2	3	4	6	5	7	9	8	10	11	12	15	13	16	17	18	14	19	20	21	22	23	24	26	25	27
1	3	2	5	4	6	8	7	9	10	11	12	13	14	16	18	17	15	19	20	21	24	23	22	25	26	27
1	2	3	4	5	6	7	8	9	10	11	12	14	13	15	17	18	16	19	20	21	22	23	24	25	27	26
1	2	4	3	5	6	9	10	8	11	7	12	14	15	17	13	18	16	19	20	21	23	22	26.5	25	26.5	24
2	1	4	3	5	6	9	10	8	11	7	12	14	16	17	13	18	15	19	20	21	23	22	24	26	26	26
1	2	3	4	5	6	8	11	9	7	10	12	13	14	17	15	18	16	19	20	21	22	26.5	23	25	26.5	24
1	2	3	4	5	6	7	11	8	10	9	12	13	14	16	15	18	17	19	20	21	22	23	24	25	27	26
1	2	3	4	5	6	7	11	8	10	9	12	14	15	17	13	18	16	19	20	21	22	23	24	25	26.5	26.5
1	2	3	4	5	6	7	11	8	10	9	12	13	15	16	14	18	17	19	20	21	22	24	23	25	26.5	26.5
1	2	3	4	5	6	7	11	8	10	9	13	15	16	17	14	18	12	20	19	21	23	22	25	24	26.5	26.5
1	2	4	3	5	6	7	11	8	10	9	12	13	15	16	14	17	18	19	20	21	23	22	24	26	26	26
1	2	4	3	5	6	7	9	8	10	11	12	13	14	17	15	18	16	19	20	21	22	23	24	25	27	26
1	2	3	4	5	6	7	10	8	9	11	12	13	14	15	17	18	16	19	20	21	22	24	23	25	26.5	26.5
1	2	3	4	6	5	7	8	9	10	11	12	14	13	15	18	17	16	19	20	21	23	22	24	25	27	26
1	3	2	4	5	6	8	7	10	9	11	12	13	14	16	17	15	18	19	20	21	23	22	24	26.5	25	26.5
1	2	3	4	5	6	7	9	8	10	11	12	14	13	15	16	17	18	19	20	21	22	23	24	25	26	27
1	2	3	6	4	5	8	7	9	11	10	12	13	14	17	16	15	18	19	20	21	23	22	24	25	26.5	26.5
1	2	3	4	6	5	7	8	10	9	11	12	14	13	15	18	16	17	19	20	21	23	22	24	25	26	27
1	2	3	6	4	5	7	8	10	9	11	12	13	14	17	16	15	18	19	20	21	23	22	24	25	27	26
1	3	2	4	5	6	8	7	10	9	11	12	13	14	16	17	15	18	19	20	21	23	22	24	25	26	27
1	3	2	4	6	5	8	7	10	9	11	12	13	14	15	17	16	18	19	20	21	23	22	24	25	26	27
1	3	2	4	6	5	7	8	9	10	11	12	13	14	16	17	15	18	19	20	21	22	23	24	25	26	27
1	3	2	4	5	6	7	9	8	10	11	12	13	14	17	16	15	18	19	20	21	23	22	24	25	26	27
1	2	3	6	4	5	9	7	8	10	11	12	14	13	16	18	15	17	19	20	21	24	22	23	25	26	27
1	2	3	4	5	6	7	9	8	10	11	12	13	14	15	18	16	17	19	20	21	23	24	22	25	26.5	26.5
1	2	3	4	5	6	8	7	9	11	10	12	13	14	18	16	15	17	19	20	21	22	23	24	26	26	26
1	2	3	4	5	6	7	9	8	10	11	12	13	14	17	18	15	16	19	20	21	22	24	23	25	26	27
1	3	2	4	5	6	7	9	8	10	11	12	13	14	17	18	15	16	19	20	23	21	24	22	25	26.5	26.5
1	2	3	4	5	6	7	9	8	10	11	12	13	14	16	18	17	15	19	20	21	23	24	22	25	26.5	26.5
1	2	3	4	6	5	7	8	9	10	11	12	14	13	16	18	15	17	19	20	21	23	22	24	25	26	27
1	3	2	6	5	4	7	8	10	9	11	12	13	14	16	17	15	18	19	20	21	23	22	25	24	27	26
1	3	2	6	4	5	8	9	7	10	11	12	13	14	17	18	15	16	19	20	22	21	23	24	25	27	26
1	2	3	6	4	5	9	10	7	8	11	12	13	14	16	18	15	17	19	20	21	23	22	24	25	27	26
1	2	3	4	6	5	7	8	9	10	11	12	13	14	15	17	18	16	19	20	21	22	23	24	25	26	27
1	2	3	4	5	6	7	8	9	10	11	12	13	14	18	17	16	15	19	20	21	23	22	24	25	26	27
1	2	3	4	5	6	7	10	8	11	9	12	15	13	14	16	18	17	19	20	21	22	23	24	26	26	26
1	2	3	4	5	6	7	9	8	10	11	12	13	14	17	16	15	18	19	20	21	23	22	25	26	24	27
1	2	4	3	5	6	8	9	10	11	7	12	14	16	17	13	18	15	19	20	21	22	23	24	25	26.5	26.5
1	2	3	5	4	6	8	11	9	10	7	12	13	16	15	14	17	18	19	20	21	22	23	25.5	25.5	25.5	25.5
1	2	3	4	5	6	7	10	8	9	11	12	13	14	15	16	18	17	19	20	21	22	23	24	25	26.5	26.5
1	2	3	5	4	6	7	8	9	10	11	12	14	13	15	18	17	16	19	20	21	23	22	26	26	24	26
1	2	3	4	5	6	7	8	9	11	10	12	13	14	15	16	17	18	19	20	21	23	22	24	25	26.5	26.5
2	1	3	4	5	6	7	9	10	11	8	12	13	14	16	17	15	18	19	20	21	23	22	24	25	26.5	26.5
2	1	3	4	5	6	8	9	11	10	7	12	13	16	15	17	14	18	19	20	21	23	22	24	26.5	25	26.5
2	1	3	4	5	6	7	8	10	11	9	12	14	17	13	15	16	18	19	20	21	23	22	25.5	25.5	25.5	25.5
2	1	3	4	5	6	7	9	10	11	8	12	14	17	13	15	16	18	19	20	21	23	22	24	25	26	27
1	3	2	5	4	6	7	8	10	9	11	12	13	15	17	16	14	18	19	20	21	22	23	24	26	25	27

(III) One can even dispense with the identity of letters and their ranks. It is sufficient if a global measure without respect to the individuality of letters can be considered constant. Out of a number of possibilities (cf. Altmann, Lehfeldt 1981) we choose two simple measures, namely the *repeat rate*

$$(5) \qquad R = \sum_{i=1}^{K} p_i^2$$

and the *entropy*

$$(6) \qquad H_1 = -\sum_{i=1}^{K} p_i \, ld \, p_i$$

with $p_i = f_i/N$, $N = \Sigma f_i$, where f_i are the absolute frequencies.

These two global measures can be compared with those in other languages. Altmann and Lehfeldt (1981) collected R and H_1 for 63 languages from the available literature. Though not all counts were reliable it could be shown that both R and H_1 depend on K, the number of entities (phonemes or letters) in the inventory. As can be seen in Altmann and Lehfeldt (1981, Figures 4.3 and 4.4, p. 158 and 173 respectively), R and H_1 for Finnish lie perfectly among the other points denoting these measures for the other languages.

The above procedures should be automatically appended to any phoneme or letter count.

Searching for laws

(IV) In quantitative linguistics - as in all other more mature empirical sciences - it is usual to assume the existence of laws behind all phenomena. If we succeed in finding them then we gain a deeper insight into the mechanisms generating these phenomena and have better chances of embedding them in linguistic control cycles which is on a par with an explanation (cf. Salmon 1984). Here we content ourselves with the first steps.

The usual way to model the rank-frequency distribution is the use of Zipf's law and Zipf-Mandelbrot's law (cf. Guiter, Arapov 1982, Orlov, Boroda, Nadarejšvili 1982). Unfortunately, the researchers always avoided any testing for goodness of fit. This was perhaps caused by the fact that the sample sizes were very large and the chi-square test had yielded significant deviations. However, one can obtain good results with small sample sizes (cf. Altmann 1988:69-77). Several authors successfully used (with small sample sizes) a distribution developed by Hammerl (esp. 1991) however not for phoneme/letter counts.

In order to avoid problems arising with large sample sizes that are always present when we work with the (statistical) distribution of phonemes or letters, we try to find the curve for the ranks of the relative frequencies or proportions. This is not a probability distribution but a sequence of values. We start from the fact that the differential equation leading to the Zipf-Mandelbrot law has the form

$$(7) \qquad \frac{dy}{y} = -\frac{c}{a+x}dx$$

In the case of ranking we have discrete steps for which $dx = 1$ and $dy = y_x - y_{x-1}$, so that we can write (7) as

$$(8) \qquad \frac{y_x - y_{x-1}}{y_{x-1}} = -\frac{c}{a+x}$$

(writing x instead of x-1 on the right side of (8) does not change the result) yielding

$$(9) \qquad y_x = \frac{a - c + x}{a + x} y_{x-1}$$

Writing $a - c = b$ (with $b < a$) and letting $y_x = 0$ for $x < 1$ we obtain the solution

$$(10) \qquad y_x = \begin{cases} y_1 & for\ x = 1 \\ \dfrac{(b+2)(b+3)\ldots(b+x)}{(a+2)(a+3)\ldots(a+x)}\, y_1 , & for\ x = 2,3,\ldots \end{cases}$$

or

$$(11)^1 \quad y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} y_1 , \qquad for\ x = 1,2,3,...$$

Since $y_x > 0$ (for x = 1,2,3,...) and it decreases monotonically, the parameters fulfill one of the conditions (i) $a > b \geq 0$, (ii) $a > 0$, $-1 < b < 0$, (iii) $b < a < 0$ with [a] = [b], a and b are not integers. Formula (9) (with a-c = b) can be used for recursive computation of the y_x-values.

As can be seen, (10) or (11) converges to the (displaced) geometric series, if $a \to \infty$, $b \to \infty$ and $b/a \to q$, since

$$(12) \quad \lim \frac{(b+2)...(b+x)}{(a+2)...(a+x)} y_1 = y_1 q^{x-1}$$

Sigurd (1968) used the geometric series as a model for the rank-frequency distribution and later on it was used as the first approximation to the models of repeat rates and entropies (cf. Altmann, Lehfeldt 1981). Since many authors only published relative frequencies, it is more reasonable to work with a curve, than with a probability distribution where the goodness-of-fit test can lead to the rejection of the model. The probabilistic counterpart of (11) is the hyper-pascal distribution with q = 1 and y_1 as the norming factor (cf. Altmann 1991). The fitting of (11) to Pääkkönen's data is shown in Table 4. The goodness of fit has been ascertained simply by means of the determination coefficient which in this form merely compares the residual squared deviations with total squared deviations:

$$(13) \quad D = 1 - \frac{\sum_{x=1}^{K} (y_x - \hat{y}_x)^2}{\sum_{x=1}^{K} (y_x - \bar{y})^2}$$

where y_x = observed values
$\hat{y}_x$ = computed values (according to 11)
$\bar{y}$ = mean of observed values (here always 100/K).

[1] If we consider (11) as a probability function then $y_1 = (b+1)(a-b-1)/a(a+1)$ is the norming constant.

The fit is the better the greater D is. Problems connected with D and with curvilinear fits are discussed by Grotjahn (1992). All computations were made with the simplex method of Nelder and Mead (1965). As can be seen all fits are quite good even if one does not expect good results with mixed samples.

The parameters a and b of (11) seem to be dependent on each other. Comparing a and b in Table 4 we can observe that for Fiction, Nonfiction, Newspapers, KH, New Testament and S3 b/a = 0.8, for S2 it is 0.83 and for Total it is 0.86. Knowing this fact we can begin the optimization by Nelder and Mead with other starting values and obtain still better results. Some examples are in Table 5. If with increasing parameters the fit improves and b/a converges to a constant then we have an indication that the geometric series would yield a good fit. This is merely the consequence of (12). The case of Hawaiian with small values of parameters taken from Pukui and Elbert (1957) indicates that if we compare this case with the other ones the geometric series (yielding D = 0.8078) is not adequate (s. Table 6).

If the above model is correct, then rank-frequency distributions must display a strong regularity even with small sample sizes. This would of course hold only if no factors would operate on phoneme/letter frequencies. Needless to say, this is not so. In every language and in every text there are a lot of boundary conditions influencing the phoneme frequencies. They can be of the following kinds: (a) *local* ones, e.g. the kind of the text, the style of the author, the situation in which the text arose, the individuality of the hearer for whom the text was generated, even the planned length of the text (according to Orlov 1982), etc. and (b) *global* ones, consisting of the influences of functional equivalents in the same control cycle and of the influences of entities lying at higher levels, e.g. all phonological and morphological means used for generating redundancy, frequencies of words, word length, etc.

Many rank-frequency distributions or curves must be fitted before we begin to set up hypotheses about the parameters of (11). On the other hand, all functional equivalents of the frequency distribution of phonemes must be found and evaluated before we gain a deeper insight into this mechanism.

In any case, the possibility of fitting the same function to all rank-frequency sequences is a strong support for the assumption that they follow a law. Thus the proportion at a given rank is stable, i.e. if phonemes/letters exchange their ranks the proportions at the given ranks remain statistically unchanged.

The consequences of this fact are, however, still deeper. If we can show that the proportions of entities at this level follow a law then the same must hold

for all other levels of language but, because of the greater size of all 'higher' entities (e.g. syllables, morphemes, words, phrases etc.), we attain convergence merely with enormous sample sizes if the individuality of the entities is taken into account. The more we neglect the individuality of entities, the more rapidly we shall observe a concordance with a law. Since from the theoretical point of view individual language units are not relevant, we have a good chance of finding rank-frequency laws for all other language phenomena. But this has been known for a long time (cf. Arapov 1989).

Table 4. Fitting of (11) to Pääkkönen's data

x	Fiction y_x	Fiction $\hat{y}_x$	Nonfiction y_x	Nonfiction $\hat{y}_x$	KH y_x	KH $\hat{y}_x$	Newspapers y_x	Newspapers $\hat{y}_x$
1	11.74	11.74	11.83	11.83	12.34	12.34	12.20	12.10
2	11.30	9.84	10.50	9.90	11.42	10.37	10.64	10.11
3	9.34	8.39	9.96	8.43	9.83	8.87	9.59	8.59
4	9.09	7.26	8.55	7.27	8.79	7.69	8.72	7.39
5	7.79	6.35	8.31	6.34	7.35	6.75	8.19	6.43
6	6.96	5.61	8.16	5.59	7.16	5.98	7.66	5.66
7	5.85	4.99	5.56	4.97	6.15	5.35	5.82	5.01
8	5.75	4.48	5.28	4.45	5.78	4.81	5.34	4.48
9	5.58	4.05	5.26	4.01	5.61	4.36	5.33	4.03
10	5.18	3.68	5.08	3.63	4.88	3.98	4.99	3.64
11	4.77	3.36	4.48	3.31	4.48	3.65	4.40	3.31
12	3.18	3.09	3.28	3.03	2.81	3.36	3.38	3.02
13	2.40	2.84	2.58	2.79	2.30	3.10	2.47	2.77
14	2.31	2.63	2.34	2.57	2.19	2.88	2.36	2.55
15	2.13	2.44	1.88	2.38	2.15	2.68	2.03	2.36
16	2.05	2.27	1.84	2.21	2.14	2.50	1.90	2.19
17	1.89	2.12	1.68	2.06	2.10	2.34	1.77	2.03
18	1.58	1.99	1.65	1.93	1.59	2.20	1.71	1.90
19	0.57	1.87	0.93	1.80	0.43	2.07	0.80	1.77
20	0.40	1.76	0.53	1.69	0.40	1.95	0.46	1.66
21	0.07	1.66	0.13	1.59			0.14	1.56
22	0.03	1.56	0.07	1.50			0.07	1.47
23	0.02	1.48	0.07	1.42			0.06	1.38
24	0.01	1.40	0.04	1.34			0.05	1.31
25			0.01	1.27			0.02	1.24
	a = 8.3974 b = 6.7179 D = 0.9184		a = 8,7831 b = 7.0265 D = 0.9489		a = 7.9063 b = 6.3250 D = 0.9475		a = 9.1863 b = 7.3490 D = 0.9601	

KH = data collected by Kaisa Häkkinen

Table 4 (Cont.)

x	New Testam. y_x	New Testam. $\hat{y}_x$	S2 y_x	S2 $\hat{y}_x$	S3 y_x	S3 $\hat{y}_x$	Total y_x	Total $\hat{y}_x$
1	11.93	11.93	11.14	11.14	11.90	11.90	11.62	11.62
2	9.95	10.00	10.83	9.55	10.64	9.96	10.71	10.07
3	9,83	8.53	10.07	8.27	9.77	8.47	9.88	8.77
4	9.31	7.38	8.68	7.24	8.67	7.30	8.67	7.67
5	9.28	6.46	8.22	6.39	8.21	6.37	8.21	6.74
6	7.30	5.70	7.87	5.68	7.85	5.61	7.86	5.95
7	5.70	5.08	5.89	5.08	5.68	4.98	5.76	5.27
8	5.70	4.56	5.43	4.57	5.34	4.45	5.31	4.68
9	5.09	4.12	5.19	4.13	5.24	4.01	5.27	4.18
10	4.69	3.75	5.16	3.76	5.06	3.63	5.00	3.74
11	4.66	3.42	4.88	3.43	4.59	3.31	4.81	3.35
12	3.29	3.14	3.86	3.14	3.30	3.03	3.51	3.02
13	2.72	2.90	2.32	2.89	2.52	2.78	2.45	2.72
14	2.62	2.68	1.96	2.67	2.32	2.57	2.16	2.46
15	2.20	2.49	1.89	2.47	1.91	2.37	1.93	2.23
16	1.59	2.32	1.85	2.29	1.83	2.21	1.82	2.03
17	1.52	2.17	1.84	2.14	1.79	2.05	1.81	1.85
18	1.18	2.03	1.53	1.99	1.74	1.92	1.66	1.69
19	0.83	1.90	0.84	1.86	0.85	1.80	0.84	1.54
20	0.29	1.79	0.45	1.75	0.49	1.69	0.47	1.42
21	0.10	1.69	0.07	1.64	0.13	1.59	0.11	1.30
22	0.06	1.60	0.03	1.55	0.06	1.49	0.05	1.20
23	0.06	1.51	0.03	1.46	0.06	1.41	0.05	1.10
24					0.04	1.33	0.03	1.02
25					0.01	1.26	0.01	0.94
	a = 8.3566 b = 6.6853 D = 0.9129		a = 12.0338 b = 10.0282 D = 0.8910		a = 8.8883 b = 7.1107 D = 0.9264		a = 28.0998 b = 24.0855 D = 0.9337	

Conclusions

(1) **Large sample** sizes have all the advantages mentioned. Global language characteristics are quite stable but in order to achieve stability enormous samples must be drawn. The size must be the greater the higher the counted linguistic unit. For words perhaps hundreds of millions of items must be counted.

Table 5. Improved fits of data in Table 4

	a	b	D	b/a	geometric	
					q	D
Fiction	27.86 188.65 515.60	23.88 166.27 455.92	0.9373 0.9589 0.9616	0.86 0.88 0.88	0.8772	0.9666
Nonfict.	15.33 30.26 5566.31	12.77 25.94 4922.10	0.9211 0.9403 0.9647	0.83 0.86 0.88	0.9027	0.9429
KH	26.08 476.34	22.35 421.19	0.9615 0.9773	0.86 0.88	0.8821	0.9792
Newsp.	16.27 33.13 8337072.04	13.56 28.40 7372182.97	0.9354 0.9530 0.9744	0.83 0.86 0.88	0.8914	0.9726
NT	28.31 1170.06	24.27 1034.60	0.9327 0.9585	0.86 0.88	0.9101	0.9097
S2	21.47 112.61	18.40 99.58	0.9070 0.9367	0.86 0.88	0.9060	0.9322
S3	30.92 2932425.94	26.50 2593040.29	0.9451 0.9686	0.86 0.88	0.8965	0.9619
Total	513.33	453.91	0.9598	0.88	0.8996	0.9529

It is questionable whether a small mixture of texts is an appropriate sample at all. Two individual samples may show an agreement with a law but their mixture can display a significant deviation so that formula (11) must be modified in such cases.

Full individual texts are populations of their own.

Testing the goodness-of-fit or homogeneity by means of the chi-square test is somewhat problematic with large samples (cf. Grotjahn, Altmann 1992). The same holds for the *t*-test for differences between individual proportions in large samples. Therefore it is perhaps better to perform some global tests on ranks and neglect the individuality of phonemes.

(2) Letter counts are necessary both for practical and theoretical purposes. In theoretical investigations they are instances for testing the adequacy of laws.

Formula (11) is an attempt to establish such a law. It has the advantage that it is linguistically interpretable as an interaction between the speaker (nominator) and the hearer (denominator), who operate linearly upon the relationship between two neighbouring frequencies (y_x,y_{x-1}) (cf. Altmann 1991). If laws could be found, then our only task would be to compare our samples with them, and if bad agreement is found then formulas like (11) must be modified taking into account the specificity of individual counts. (This is the reason why mixtures of counts should be avoided.) For the time being formula (11) unifies the existing approaches (Sigurd's and Mandelbrot's), but new evidence may force us to make modifications.

(3) When publishing phoneme/letter/word etc. counts one should always give both absolute and relative frequencies and present these numbers for all individual samples (texts) separately. Their summation is merely a supplement. The data should be at least commented on in English.

(4) The problems to be solved are: (a) Testing formula (11) on many counts and (b) finding the factors that are responsible for the realized values a and y_1, i.e. to embed (11) in an appropriate control cycle.

Table 6. Fitting (11) and geometric series to Hawaiian

	(11)		geometric	
x	y_x	$\hat{y}_x$	y_x	$\hat{y}_x$
1	26.45	26.45	26.45	26.45
2	11.94	15.69	11.94	19.52
3	11.13	11.13	11.13	14.41
4	9.34	8.62	9.34	10.64
5	8.73	7.03	8.73	7.85
6	6.98	5.93	6.98	5.80
7	6.34	5.13	6.34	4.28
8	5.80	4.52	5.80	3.16
9	5.35	4.04	5.35	2.33
10	5.01	3.65	5.01	1.72
11	1.96	3.32	1.96	1.27
12	0.55	3.05	0.55	0.94
13	0.40	2.82	0.40	0.69
	a = 0.4910 b = -0.5225 D = 0.9274		q = 0.7382 y_1 = 26.45 D = 0.8078	

References

Altmann, G. (1987). The levels of linguistic investigation. *Theoretical Linguistics 14, 227-239.*

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

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification processes in language: grammar.* Hagen, Rottmann: 33-46.

Altmann, G., Lehfeldt, W. (1981). *Einführung in die quantitative Phonologie.* Bochum, Brockmeyer.

Arapov, M.V. (1989). *Matematičeskaja lingvistika.* Moskva, Nauka.

Grotjahn, R. (1992). Evaluating the adequacy of regression models: Some potential pitfalls. *Glottometrika 13, 121-172.*

Grotjahn, R., Altmann, G. (1992). Modelling the distribution of word length: some methodological problems. *Proceedings of the First Quantitative Linguistic Conference (QUALICO) in Trier, Sept. 23-27, 1991 (to appear)*

Guiter, H., Arapov, M.V. (eds.) (1982). *Studies on Zipf's law.* Bochum, Brockmeyer .

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier, WVT.

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

Nelder, J.A., Mead, R. (1965). A simplex method for function minimization. *Computer Journal 7, 308-313.*

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

Pukui, H.K., Elbert, S.H. (1957). *Hawaiian-English dictionary.* Honolulu, University of Hawaii Press.

Salmon, W.C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, N.J., Princeton UP.

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

Glottometrika 14, 1993, 69-81

Measuring Text Difficulty [1]

Juhan Tuldava, Tartu

In recent years, the measuring of text difficulty (readability) has become ever increasingly topical not only in didactics and psychology but also in philology and stylistics. Objective, quantitative methods of measuring text difficulty are of great importance in determining the level of educational text comprehensibility and adequacy, readability of the press, and clear writing of scientific and popular-scientific papers.

How could text difficulty be defined? This paper distinguishes subjective and objective difficulty. Subjective difficulty, or the reader's difficulty, can be measured in different ways: through the time spent on reading the text, answering questions about the text, expert opinions and other methods. Objective difficulty (text complexity) is expressed through the formal and semantic textual characteristics which correlate with the subjective assessments of text difficulty and, therefore, can be used as prognostic characteristics of the subjective difficulty of the text. By combining subjective assessment and objective quantitative characteristics scholars have worked out a number of formulae of readability. Readability formulae have been successfully used in education and psychology in a number of countries for dozens of years. Without being too detailed we shall mention some of the approaches in this field (for detailed discussion see, e.g., Klare 1974-75).

The readability formula by R. Flesch (1948) is very popular. The formula includes the formal characteristics of the mean sentence length, the mean word length, and correlates them with experimentally found subjective characteristics of text difficulty. The formula is expressed as an equation of multiple linear dependence:

$$R.E. = 206.835 - 0.846\ wl - 1.015\ sl, \qquad (1)$$

where R.E. = Reading Ease (experimental readability assessment), wl = the number of syllables per 100 words in the text, sl = the mean sentence length in words. This formula has found wide application in the assessment of readability of English texts.

[1] Revised and extended version of an article (Tuldava 1975) translated from Russian.

A very simple method of text readability measurement has been in use in Sweden. The formula (Sigurd 1970:137) is

$$I = L_m + L_0 , \quad (2)$$

where I is the index of text difficulty (readability), L_m is the mean sentence length in words (in a sample of 200 sentences) and L_0 is the percent of words over six characters (in a sample of 2,000 words). The formula applied to Swedish school textbooks resulted in the following assessment:

Form	1	2	3	4	5	6	7	8	9
index	18	22	26	29	31	33	35	37	39

It can be seen that the index values are higher for the textbooks of higher forms as could be expected. It is of interest that the editors of some Swedish newspapers have used readability formulae to see that the published material is of adequate difficulty to meet the requirements of the average reader. Texts with an index over 50 are considered excessively difficult (Sigurd 1970:138).

A noteworthy contribution to the theory and practice of measuring text difficulty and readability has been made by the Estonian scholar Jaan Mikk (1974; 1991), who, on the basis of extensive experimental evidence and with the use of correlation and factor analyses, has established the most important characteristics of subjective text difficulty and the corresponding formal prognostic characteristics of text. Mikk has developed several formulae of readability for texts in Estonian, English and Russian. The most time-saving and economical readability formula for Estonian is the following:

$$C = 0.131\ X + 9.84\ Z - 4.6,$$

where C is the index of readability correlating with subjective difficulty assessment (the percent of errors in the cloze test of the experiment), X is the mean length of the self-contained sentences in characters (a sentence is considered to be "self-contained", if it is a simple sentence or a clause in the complex or compound sentence), Z is the mean abstractness of the repeating nouns (i.e. occurring more than once) in the text, measured in a special way suggested by Mikk (1974:118).

As shown by Mikk's research the most reliable ways of measuring the subjective level of text difficulty are 1) expert opinion and 2) the percent of errors in the cloze test (in the experiment with Estonian texts each seventh word had been deleted). These two factors were the most important ones and became

Factor 1 in the factor analysis which included as other factors such experimental characteristics as the knowledge increase caused by the reading of the text, the adequacy of the conclusions drawn, the number of misinterpreted phrases and others (Mikk 1974:132). Of the formal textual characteristics the most reliable ones in text difficulty prognostication were the sentence length, the word length and the characteristics of word abstractness including the percentage of abstract nouns with abstract suffixes such as *-us, -mine, -ism* (correspondingly *-ity (-ness), -ing, -ism* in English). Significant correlations were also established between the subjective level of text difficulty and textual characteristics such as the level of phrasal difficulty (measured by the number of interword connections), the percentage of the international words in the text, and the percentage of the familiar words in the text (the list of familiar words in the text was drawn up with the help of a frequency dictionary).

To measure the prognostic qualities of the objective level of text difficulty through textual characteristics in the ideal variant, all the formal textual characteristics correlating with the subjective assessment of text difficulty should be studied. But in practice, studies of text difficulty measurement are governed by the rules of sparing use of time and effort. To simplify the measurement and still be representative and valid, certain formal textual characteristics can be selected as they are known to correlate well with one another. We could see that formulae of readability usually contain only a limited number of textual characteristics, which are correlated with the results of the experimental assessment of the subjective level of text difficulty. The shortcoming of such readability formulae lies in the fact that the values of subjective assessment tend to vary and, consequently, the corresponding constants in the formulae are only of relative adequacy. The question arises of a possible formula with a more stable assessment level. This formula should be based on the research done so far, and it should proceed to select analytically the characteristics of prognosis most suited for the purpose and show their mutual relations in text generation. We should like to draw the reader's attention to a possible formula of that kind.

To build up a formula for measuring the objective level of text difficulty, one must first decide upon the selection of adequate characteristics of good prognosticating power. Naturally, one will draw on the work done before. Analyzing readability formulae, it is of interest to mention that nearly all of them contain the argument of sentence length. However, they interpret the term "sentence" in different ways. Mikk in his readability formulae includes "the length of the self-contained sentences". This approach requires previous syntactic analysis of the sentences in the text. We think that one could proceed from the "completed sentences", i.e. the traditional interpretation of the sentence, ignoring the structural peculiarities. Besides, the mean values of the self-contained sentence

length and the completed sentence length are highly correlated (according to Mikk, the correlation coefficient is +0.92). It is thought that the reader's idea of syntactic complexity, easiness/cumbersomeness of text, is very much dependent on the length of the completed sentence in it (Akimova 1973).

The length of a sentence can be measured in different ways. If the sentence length is measured in characters including letters, punctuation marks and word spaces (see Mikk 1974:116), then the measurement indirectly includes some information about the word length. In a simpler and also more adequate way the sentence length is measured in words. By automated measurement a word is any sequence of letters between two word spaces. The sentence length measured in words can easily be correlated with other characteristics of syntactic complexity of the sentence, such as sentence depth (Martynenko 1971) or the percentage of composite sentences in the text (Akimova 1973:70). There is no doubt that the mean sentence length in words reflects the logical organization of thought and its level of complexity well, and, therefore, appears as one of the most important characteristics of the objective level of text difficulty, prognosticating the difficulty level of the text for the reader. This conclusion is suggested by a number of studies on various European languages. (In this connection, we have studied texts in English, German, Swedish, Estonian, Russian and Bulgarian.)

So one of the diagnostic textual characteristics must be the mean sentence length in words.

To find another component for the formula is not so easy. Many studies emphasize the importance of word length in text comprehension. Longer words are considered to be more complex and more informative than shorter ones. There is a negative correlation between the word length and its frequency of occurrence, i.e. longer words are usually not as frequent as shorter ones, and consequently they are also less familiar and more difficult to comprehend. When word length was analyzed in its correlation with the frequency rank in the frequency dictionary of Swedish in the corpus of the first 1,000 most frequent words, the coefficient of rank correlation was -0.80. The results of the text study by Mikk (1974) demonstrate good correlations between mean word length and the abstractness level of the nouns in the text, the percentage of the international words in the text and the complexity level of the phrases, which all in good probability prognosticate certain aspects of text difficulty for the reader. However, the sentence length and the word length also correlate well, which allows Mikk to discard the mean word length as an independent characteristic of text difficulty and include it in the group of sentence length characteristics. And yet careful analysis has shown that the mean word length correlates better

with a number of important characteristics of text difficulty than the mean sentence length. Quoting Mikk, the correlation of mean word length with the number of international words in the text is +0.61, while that of the mean sentence length is +0.29; the mean word length correlation with the percentage of abstract nouns in the text is +0.55, while that of the mean sentence length is +0.33; the corresponding figures for the correlation with the percentage of familiar words in the text are -0.80 and -0.49 respectively. Here one must note that in the calculations made above, sentence length and word length are measured in characters, i.e. word length is indirectly reflected in sentence length. So it can be said that the mean sentence length and the mean word length reflect both similar and different aspects of text difficulty.

It should not be neglected that the mean word length has an independent role to play in prognosticating the text difficulty level. There is no denying that the characteristic of the mean word length can be replaced by other reliable characteristics measuring text difficulty, such as the percentage of abstract nouns in the text or the percentage of familiar words in the text, but it is to be remembered that the mean word length correlates well with the characteristics mentioned above, and it is the characteristic easiest to subject to automated text measurement.

The word length can be measured in different units (characters, letters, phonemes, syllables, morphemes). By automated measurement it can be done most easily in characters or letters. Anyhow all ways of measurement should correlate with one another. We have chosen for our purpose the measurement of word length in syllables, as this is the way it has been measured in the studies the data of which we are going to use. Syllable counting seems to be the most convenient way of non-automated data processing. Besides, measuring the mean word length in syllables is closest to the measurement used to establish the word depth (i.e. word length measured in morphemes) characterizing the morphological structure of the word. There is a direct correlation between the word depth and the word length measured in syllables. So *the mean word length measured in syllables* suits our purposes as one of the diagnostic characteristics of text difficulty.

The analysis of the works by R. Flesch, J. Mikk and others shows that the formulae of text difficulty (readability) can be built up on two factor characteristics. So we have chosen the mean sentence length and the mean word length for our study, as they cover the essentials of the objective text difficulty level sufficiently well and each of them characterizes a different aspect of the subjective text difficulty level. The correlation between the two reflects the subject matter which is jointly covered by the two characteristics. Now the next step

would be to decide upon the form of the formula. R. Flesch and others in their formulae correlate the textual characteristics with the subjective assessment of text difficulty. As we have already noted, the outcome of that approach fully depends on the way the subjective measurements are conducted. So we shall seek to find another, more objective way of measuring the interrelationship of the components of objective text difficulty.

In a less known study of W. Fucks (1956), the quantitative stylistic characterization of text has been measured by the formula:

$$\bar{i}^{\alpha}\,\bar{j}^{\beta} \qquad (3)$$

where $\bar{i}$ is the mean word length in syllables and $\bar{j}$ is the mean sentence length in words, α and β are constants depending on the relative significance of i and j seen from the point of view of text difficulty assessment. Fucks suggests a simplified version of $\alpha = \beta = 1$ and then the formula will become $i.j$, i.e. the mean word length is multiplied by the mean sentence length.

As the characteristics of the mean word length and the mean sentence length prognosticate the subjective level of text difficulty well, the Fucks` formula might be used in measuring text difficulty. But if the equation $\alpha = \beta = 1$ is used, the changes in the sentence length will too strongly influence the overall outcome of text difficulty measurement. Surely the sentence length variations can be expressed in units that will proportionately correspond to the mean word length variations in the language under study. So for Russian texts the modified formula in which $\alpha = 1$ and $\beta = 0.2$ was found to be most appropriate.

And yet we decided to continue our search for some more reliable foundation for our formula of text difficulty measurement. After we had examined the relationship between the mean sentence length and the mean word length in the texts of several languages, we could establish an approximate linear relationship between the word length and the logarithm of the mean sentence length. The regularity is most clearly seen when texts of different genres or sublanguages are compared. Graphically the regularity has been presented on Figure 1. All that means that the interrelationship between the sentence length and the word length is governed by some shared regularity. The dependence of the mean word length on the mean sentence length is an example of the law of logarithmic growth:

$$\bar{i} = a + b\ ln\,\bar{j} \qquad (4)$$

in which $\bar{i}$ is the mean word length in syllables, $\bar{j}$ is the mean sentence length in words, a and b are constants, and ln is the natural logarithm. An inverse relationship should also be there: a dependence of the mean sentence length on the mean word length can be analytically expressed as an exponential function:

$$\bar{j} = Ae^{B\bar{i}} \qquad (5)$$

where A, B are constants and e is the base of the natural logarithms.

Having established the linear relationship between the mean word length $\bar{i}$ and the logarithm of the mean sentence length $\bar{j}$, one is apt to think that the balanced linear dependence of values may serve successfully as the natural measure of the objective level of text difficulty. The formula will be written:

$$R = \bar{i}\ ln\,\bar{j} \qquad (6)$$

where R is the index of the objective level of text difficulty. Instead of ln the decimal lg can be used (as in the original version: Tuldava 1975), which will result in different numerical values, but it will not affect the relative values.

We shall give an example to illustrate the point. We used the data presented in reports on German texts (Fischer 1965; Fucks, Lauter 1965; Thiele 1968). We included the values of the mean word length and the mean sentence length of 20 German texts in our study (the list of the texts can be found at the end of the present paper). The texts belonged to different genres and sublanguages. Figure 1 presents the graph of the relationship between the mean word length and the mean sentence length in the texts. We used formula (4), in which the constants a and b were calculated with the least-squares method. The linear regression in our example assumed the expression of:

$$\bar{i} = 1.04 + 0.24\bar{j}.$$

The coefficient of linear correlation is equal to +0.55, with the critical value at 0.44 at a significance level of 0.05 and degrees of freedom 20 - 2 = 18.

The confidence limits of the regression can also be calculated (see, e.g., Draper, Smith 1968, Chapter 1.3). The dotted lines on Figure 1 represent the variation range measured at significance level 0.01, the numbered rings represent the observed textual data, and the solid line represents the regression line. Text No. 4 can be seen remaining quite near the variation range, as this newspaper text has longer mean word length than the other texts of the zone (see Table 1) - so also Texts Nos. 14, 15, 17, 20. On the other hand, two texts by Goethe (Nos.

7 and 12) and Text No. 8 (Eichendorff) also remain outside the variation range as their mean word length is shorter than could be expected judging by the values of the mean sentence length. This analysis may prove very helpful for stylistic text analysis (cf. Arens 1965; Fucks, Lauter 1965).

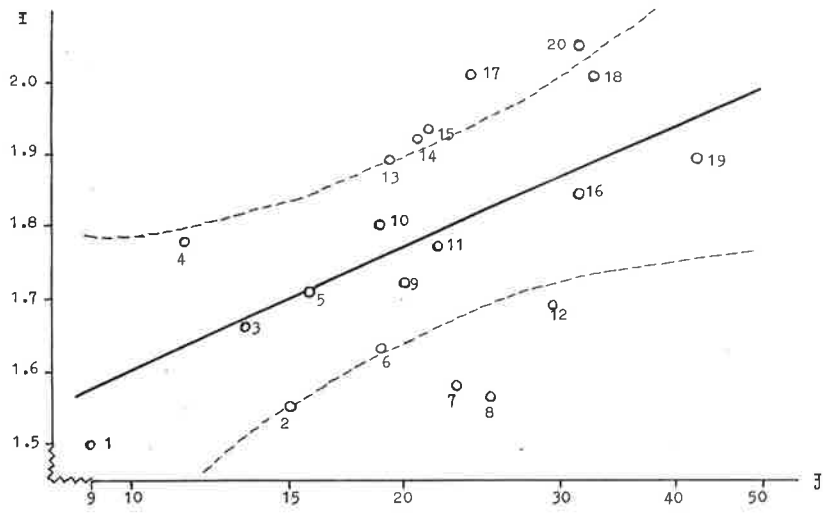


Figure 1. A correlation graph between word length in syllables ($\bar{i}$) and sentence length in words ($\bar{j}$). The abscissa axis presents a logarithmic scale. The material is in German (see Table 1).

The regression equation or its graph can be used to interpolate, i.e. to prognosticate, the expected mean word length from the mean sentence length. That makes it possible to calculate certain average theoretical values at which the empirical values should converge. For instance, one can calculate the expected word length ($\bar{i}$) in the texts where the mean sentence length ($\bar{j}$) is 20 words: $i = 1.04 + 0.24 \ln 20 = 1.76$ syllables. One should note here, however, that extrapolation exceeding the experimental data range should be avoided (in the example the mean sentence length ranged from 9 to 42 words).

The regression equation for the dependence of the sentence length on the word length can also be formed. For the given example it will equal $\ln \bar{j} = 0.88 + 1.21\bar{i}$ and $\bar{j} = 2.42e^{1.21\bar{i}}$.

Let us proceed to calculating the text difficulty index. Table 1 presents the initial data. The texts have been grouped into three subgroups of *varia* (school

textbooks, student compositions, newspaper texts), texts of *fiction* and texts of *scientific writing*. It can be seen that the index is of good stylistically discriminating quality. The texts of Group 1 (Varia) are quite clearly distinct from the texts of Group 2. The texts of Group 2 smoothly merge with the texts of Group 3. So Text 12 by Goethe, which is of a philosophical nature, seems to be more difficult than the easiest text of scientific writing.[2]

Table 1. Data of German texts: mean word length in syllables ($\bar{i}$) and mean sentence length in words ($\bar{j}$). Calculating the index of text difficulty (R) and comparing it with the readability index by Flesch (R.E.)

Genre and text (author)	$\bar{i}$	$\bar{j}$	$\ln \bar{j}$	$R = \bar{i} \ln \bar{j}$	R.E.
I. Varia					
1. Text-book (form 2)	1.50	9.0	2.1972	3.30	70.8
2. Text-book (forms 3-4)	1.55	15.0	2.7081	4.20	60.5
3. Composition by 13-year-old pupils	1.66	13.3	2.5878	4.30	52.9
4. BILD/Hamburg (1968)	1.78	11.4	2.4336	4.33	44.7
II. Texts of Fiction					
5. F. Schiller	1.71	15.6	2.7473	4.70	46.3
6. Th. Storm	1.63	18.8	2.9339	4.78	49.9
7. J.W. Goethe (1)	1.58	22.8	3.1268	4.94	50.0
8. J. Eichendorff	1.56	24.9	3.2149	5.02	49.6
9. H. Hesse	1.72	20.0	2.9957	5.15	41.0
10. Th. Mann	1.80	18.9	2.9392	5.29	35.4
11. J.Wassermann	1.77	21.4	3.0634	5.42	35.4
12. J.W. Goethe (2)	1.69	29.1	3.3707	5.70	34.3
III. Scientific Writing					
13. S. Freud	1.89	19.1	2.9497	5.57	27.6
14. W. Heisenberg	1.92	20.5	3.0204	5.80	23.6
15. A. Einstein	1.93	21.1	3.0493	5.89	22.1
16. G. Hegel	1.84	31.4	3.4468	6.34	19.3
17. M. Plack	2.02	23.5	3.1570	6.38	12.1
18. K. Marx	2.02	32.7	3.4874	7.04	2.8
19. H. Schliemann	1.89	42.1	3.7400	7.07	4.2
20. A.v. Humboldt	2.05	31.7	3.4563	7.09	1.2
Mean value	1.78	22.1	-	5.42	-
Standard deviation	0.17	8.1	-	1.03	-

[2] About statistical testing concerning the index R, see Tuldava (1993).

On the whole, the calculated difficulty indices coincide with the intuitive estimations of experienced readers. The low index of the newspaper text is a little unexpected. The sample consists of ten editorials from the German newspaper "BILD-Hamburg". The difficulty level of the texts is more or less the same as that of school compositions of 13-year-old pupils. This may reflect the editors' striving for meeting the requirements of the average reader. The easiness of the newspaper texts can be accounted for by the short sentences with long words in them. This example illustrates the fact that the objective difficulty index should be treated as an orienting characteristic in prognosticating the subjective text difficulty level through interpretation of qualitative character.

To facilitate the calculation of text difficulty we would suggest a nomogram (Figure 2). The Figure shows three scales. Let us assume that the i and j values are given. Then the points are found on the side scales and they are connected with a line (in practice it can be done with a ruler). The line crosses the Scale R at a point which represents the corresponding R value. For example: the word length is 2 syllables ($i = 2$) and the mean sentence length is 20 words ($j = 20$). Then it can easily be found that $R \approx 6$. The exact result can be found by calculating $R = 2 \ln 20 = 2(2.9957) = 5.99$. Nomograms are good when not very exact calculations are to be made quickly. It can also be used as an easy way of checking theoretical calculations.

The quality analysis of the measurement of text difficulty of these *German* texts makes it possible to distinguish certain tentative difficulty zones. For example:

$R > 7.0$	difficult text;
$7.0 \geq R > 6.0$	text above the average level of difficulty;
$6.0 \geq R \geq 5.0$	text of the average difficulty level;
$R < 5.0$	easy text.

It is clear that these difficulty zones are not universal. The scheme presented is adequate for students of philology.

The comparison of the results of our analysis with those done from formula (1) of Flesch reveals that they are very close. The corresponding rank correlation coefficient is $\varphi = -0.97$ (the correlation is negative as the formula we propose is concerned with the easiness of text).

The proposed means of text difficulty measurement is convenient, for its arguments are purely formal text characteristics - the word length and the sentence length. These characteristics can be measured without resorting to any content analysis. The procedure can be automated easily. The formal characteristics have

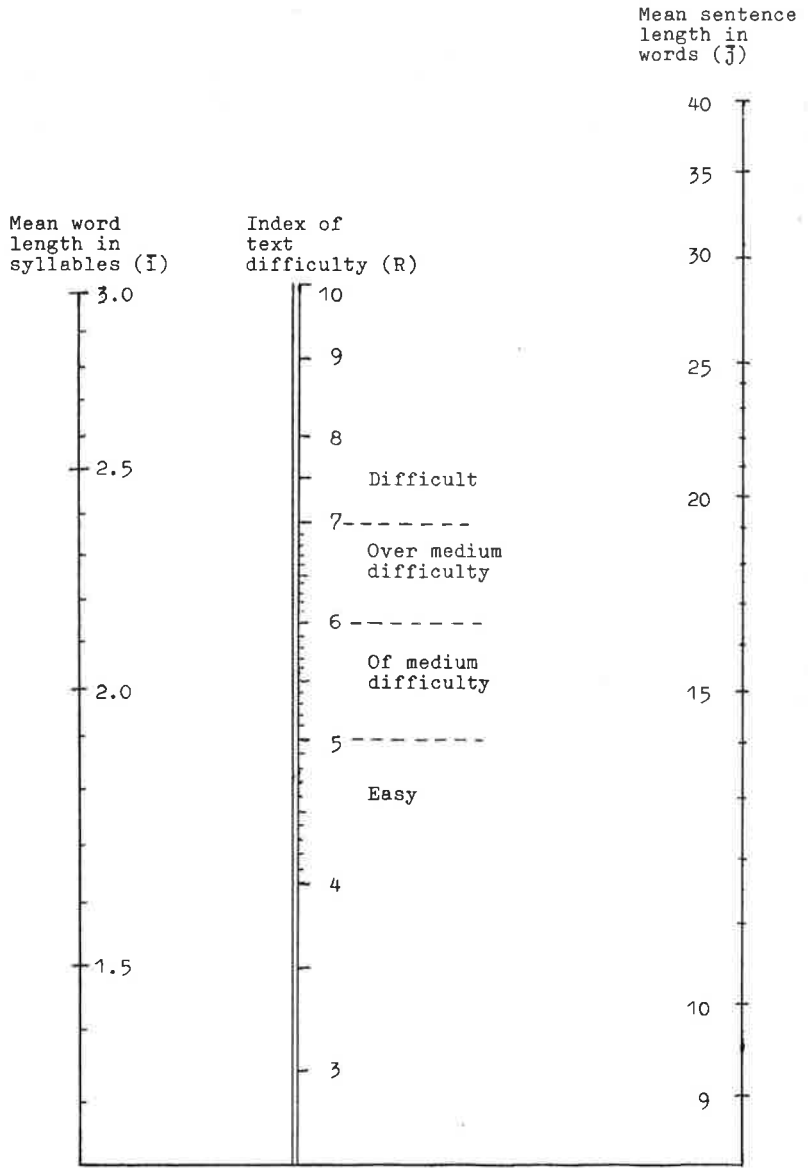


Figure 2. Nomogram for calculating the index of text difficulty (based on the formula $R = i \ln j$

been chosen to reflect rather precisely the probable level of text difficulty for the reader. The proposed quantitative index of text difficulty (formula 6) can be successfully used to distinguish texts of different individual or functional styles. The formula application has been tested on German, Estonian and Russian texts. It can be assumed to be applicable in the analysis of texts written in many other languages as well. The formula, strictly speaking, reflects a general regularity of the increment of the values of stylistic and content difficulty of texts with the growth of the size of their formal constituent textual units (on average). The probabilistic principle follows: the longer the text unit, the more complicated it is (more structured, less frequent, less familiar etc.).

In practical use, the mean word length and the mean sentence length should be measured in keeping with the requirements of the quantitative sampling method. If no great precision is required, the representative word sample will be 1,000 words, and the representative sentence sample will be 200 sentences. The samples can in their turn be divided into subsamples (e.g. 10 x 100 words and 10 x 20 sentences).

List of German texts studied
(Fischer 1965; Fucks, Lauter 1965; Thiele 1969)

1. Lesebuch für das 2. Schuljahr
2. Lesebuch für das 3./4. Schuljahr
3. Aufsätze 13-jähriger Schüler
4. 10 Leitartikel (Zeitung, BILD/Hamburg, April 1968)
5. F. Schiller. Der Geisterseher
6. Th. Storm. Der Schimmelreiter
7. J.W. Goethe. Hermann und Dorothea
8. J. Eichendorff. Aus dem Leben eines Taugenichts
9. H. Hesse. Der Steppenwolf
10. Th. Mann. Buddenbrooks
11. J. Wassermann. Der Fall Maurizius
12. J.W. Goethe. Dichtung und Wahrheit
13. S. Freud. Fragen der Laienanalyse
14. W. Heisenberg. Die Physik der Atomkerne
15. A. Einstein. Evolution der Physik
16. G. Hegel. Wissenschaft der Logik
17. M. Planck. Vorträge
18. K. Marx. Das Kapital
19. H. Schliemann. Trojanische Altertümer
20. A.v. Humboldt. Neuspanien

References

Akimova, G.N. (1973). Razmer predloženija kak faktor stilistiki i grammatiki (Sentence size as a stylistic and grammatical factor). *Voprosy jazykoznanija No. 2.*

Arens, H. (1965). *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute.* Düsseldorf, Schwann.

Draper, N.R., Smith, H. (1968). *Applied Regression Analysis.* New York-London-Sydney, Wiley.

Fischer, H. (1965). Entwicklung und Beurteilung des Stils. In: Kreuzer, H. (Hrsg.). *Mathematik und Dichtung.* München.

Flesch, R.F. (1948). A new readability yardstick. *Journal of Applied Psychology 32, 221-233.*

Fucks, W. (1956). *Zur Deutung einfachster mathematischer Sprachstatistiken.* Forschungsberichte des Wirtschafts- und Verkehrsministeriums Nordrhein-Westfalen 344. Köln und Opladen.

Fucks, W., Lauter, J. (1965). Mathematische Analyse des literarischen Stils. In: Kreuzer, H. (Hrsg.). *Mathematik und Dichtung.* München.

Klare, G.R. (1974-1975). Assessing Readability. *Reading Research Quarterly, 10, 62-102.*

Martynenko, G.J. (1971). Statističeskoe issledovanije sintaksičeskoj složnosti predloženija (Statistical analysis of sentence syntactic complexity). *Informacionnye voprosy semiotiki, lingvistiki i avtomatičeskogo perevoda 1, 84-101.*

Mikk, J. (1974). Metodika razrabotki formul čitabel'nosti (Methods of eleborating readability formulae). *Sovetskaja pedagogika i škola (Tartu) 9, 78-163.*

Mikk, J. (1991). Studies on teaching material readability. *Acta et Commentationes Universitatis Tartuensis 926, 34-50.*

Mikk, J., Mikk, E., Tirmaste, J. (1991). Computerized readability analysis of textbooks of English. *Acta et Commentationes Universitatis Tartuensis 926, 112-121.*

Sigurd, B. (1970). *Sprakstruktur. Den moderna sprakforskningens metoder och problemställningar.* Stockholm.

Thiele, J. (1969). Einige formale Daten der Ausgaben von BILD/Hamburg im April 1968. *Muttersprache 79.*

Tuldava, J. (1975). Ob izmerenii trudnosti teksta (On measuring text difficulty). *Acta et Commentationes Universitatis Tartuensis 345, 102-120.*

Tuldava, J. (1993). The statistical structure of a text and its readability. In: L. Hřebíček, G. Altmann (eds.), *Quantitative Text Analysis* (in print).

Glottometrika 14, 1993, 82-103

The Prognosticating Effectivity of Learning a Text in Physics

Hasso Kukemelk & Jaan Mikk, Tartu

The effectivity of the studying process at school depends largely on the quality of the textbook. Low quality textbooks must not be used at school even for experimental studies because it leads to incorrect view of the world. That is one of the reasons why it is important to prognosticate the effectivity of a textbook before using it at school.

The aim of this research is to elaborate learning effectivity formulas for present-days physics textbooks in Russian. Effectivity of a text includes such units as students` achievement after the learning process and their interest in the text.

The Research Methods

For this research we have taken 40 paragraphs in the physics textbooks, that have been used for some years in the 9th and 10th grades at our compulsory secondary schools (Buchovcev, Klimontovič, Mjakišev 1984; Mjakišev, Buchovcev 1985).

There are four texts that have been taken out of other textbooks for having a greater difference of readability. The paragraphs have been chosen in the second half of the physics textbook for the 9th grade and in the first half of the textbook for the 10th grade according to the table of random numbers. In this research only the basic text of a paragraph has been used.

The experimental determination of the effectivity of the texts was carried out at 13 Russian secondary schools in Estonia[1]. This research was carried out at schools by the physics teachers who had been given an additional physics lesson per week.

[1] G. Aleksina and S. Kvitko participated in the realization of the experiment

The students had to study the paragraph of the physics textbook independently. At the beginning of a lesson in order to measure the students` knowledge before the learning process they had to pass the test (one of the eight possible tests on the material of the paragraph) during which they were not allowed to use their physics textbooks or asking for the teacher's help. After doing the test (the teacher collected it) the students opened their textbooks and independently studied the paragraph during 15-20 minutes. They were not allowed to ask the teacher or other students for anything. After shutting their textbooks they filled in a form, where they wrote their opinion of how difficult and interesting the text was. Then the teacher gave the students a new test. Thus we have measured the students' achievements before and after the learning process.

In this experiment 427 students participated but for different reasons not all of them studied all the 40 paragraphs. In the following part we use the results of 304 students. They studied most of the 40 texts. There are some texts that less than 200 students studied.

The number of high-achievement-level students and low-achievement-level students influences the achieved result of the study material effectivity. It is correct to compare the material effectivity if all the texts have been learned by the students with the same ability, preparation, interest in physics, etc.

The possible influence of the difference in the students' abilities has been considered in the following way. We have calculated the average value of the students' points of tests for every paragraph. That average value shows the students' abilities, their preparation for the study process and their interest in the studies of physics. The arithmetical mean of the students' achievement (for 304 students) was 53 per cent and the standard deviation - 10 per cent. That value has been taken as the average value of the students' abilities. After that the difference of every student from that average ability was calculated. That difference was subtracted from the students' achievement of every paragraph. Thus we achieved the situation as if every student were of the average ability. Then it was possible to get the results of the students' paragraph learning even if the students' abilities were not equal to the average ability.

The second available source of deviation of the effectivity index of paragraph learning lies in the fact that the test variants on the paragraph were not equal. It is possible that students got more difficult test variants on one paragraph and more easier test variants on another paragraph. In that case it is impossible to compare the learning effectivity indexes on various paragraphs. For avoiding that situation we calculated the arithmetical mean of all final test variants of all the paragraphs. Then we calculated the arithmetical mean of the average results

of the test variants for every paragraph. In doing that we did not pay attention to the number of students who had done different variants. The arithmetical mean of the different test variants' results was taken as index of the difficulty for that paragraph. In this way we avoided possible deviations from the unequal distribution of the test variants.

The tasks of test variants were put together by G. Karu and H. Kukemelk. Statistical substantial differences between their tasks were not found. The percentage of the average right answers to the different authors' tests was equal. The paragraphs were also alike by their characteristics important for the prognosis of the study effectivity. Consequently G. Karu and H. Kukemelk had made up equal tasks and questions.

Three effectivity indexes were calculated for every paragraph according to the experiment results:

a) the student's average final achievement level in percentage;

b) the average index of interest in the paragraph. After learning the paragraph the students answered the question if that paragraph was interesting for them (2 points) or not (1 point). The arithmetical mean of that index shows the interest in the paragraph of all the students who took part in the experiment;

c) the summary index of the paragraph effectivity.

That index was found by the following formula:

$$E = \frac{L - \bar{L}}{S_L} + \frac{I - \bar{I}}{S_I} \qquad (1)$$

where E - the paragraph effectivity index;
L - the students' achievement level in that paragraph;
$\bar{L}$ - the average value of L of all paragraphs;
S_L - the standard deviation of L;
I - the index of interest in the paragraph;
$\bar{I}$ - the average I of all the paragraphs;
S_I - the standard deviation of I.

The formula (1) is made up of the standardized indexes of the students' achievement and paragraph interest (Glass, Stanley 1970). As the arithmetical mean of

the standardized quantities is zero, then the average paragraph effectivity is characterized with zero. The negative index of the effectivity shows that the effectivity of that paragraph is smaller than that of the average paragraph. The standard deviation is equal to 1 for standardized quantities. It means that they are of the same kind and so we have the same influence on the index of the paragraph effectivity as regards the text interest and the level of the students' achievement.

We analyze the basic texts in the following parts. The aim of the analysis has been to fix such text characteristics that may influence the paragraph learning effectivity. In general there are the following groups of the text characteristics:

1. The vocabulary knowledge: the larger number of unknown words in the text makes it more difficult for the students to understand and reduces the learning effectivity.

2. The abstract character of substance of the text being studied: more abstract texts are more difficult to connect with the students' experiences and so these texts are more difficult to understand.

3. The length of sentences: here difficulties may arise in connecting several parts in long sentences by students. That in its turn may lead to fragmentary understanding of sentences.

4. The complication of the text structure: inversion, long distance between connected elements, etc. break up the logical presentation.

In this article we won't describe why and how the different paragraph characteristics were found. We are limited to their computer list (table 1). Table 1 presents the arithmetical mean and the standard deviation of quantities for describing better our basic texts and textbooks.

The paragraph analysis was carried out on three methods. First of all with the help of paper and pencil we analyzed the text on the basis of Russian grammar. So we determined the values of characteristics[2] No 3-121, 145-147, 252, 436 and 448. The second method was realized with the help of the computer. All the nouns in all the paragraphs were typed in the computer in their initial forms. Together with nouns, we put in their occurrence in different language frequency dictionaries. So were determined the values of characteristics[3] No 237-260 (ex-

[2] T. Borovskaja, U. Volmer, O. Orser, J. Sivenkova, L. Tomas participated in this work.

[3] I. Sozin, E. Mikk, T. Borovskaja participated in this work.

Table 1
The list of paragraph characteristics[4]

No	The characteristics of paragraph	Arithmetical mean	Standard deviation
3	The number of printed signs	4000	1300
4	The number of words (including symbols, abbreviations)	500	170
5	The number of sentences	38	13
25	The total number of nouns	180	59
26	The total number of recurrent nouns	29	9
31	The number of illustrations in a paragraph	2.6	1.7
32	The number of formulas in a paragraph	2.7	3.6
38	The average paragraph noun occurrence in the language	198	54
39	The average paragraph noun occurrence in the physics textbooks	400	160
40	The average abstractness of nouns in a paragraph	2.0	0.2
41	The average paragraph noun occurrence of the physics textbook for grade 6	30	22
42	The average paragraph noun occurrence of the physics textbook for grade 7	55	34
43	The average paragraph noun occurrence of the physics textbook for grade 8	83	59
44	The average paragraph noun occurrence of the physics textbook for grade 9	122	65
45	The average paragraph noun occurrence of the physics textbook for grade 10	135	133
46	The average paragraph scientific-technical noun occurrence	518	90
47	The percentage of 9-and-more-letter nouns	27	8
48	The percentage of 10-and-more-letter nouns	16	8
49	The percentage of 11-and-more-letter nouns	11	7
50	The percentage of 12-and-more-letter nouns	6.5	5.6
51	The percentage of 13-and-more-letter nouns	4.5	4.4
52	The percentage of 14-and-more-letter nouns	1.9	2.9
53	The percentage of 15-and-more-letter nouns	0.8	1.2
54	The percentage of nouns that are not in the spoken language frequency dictionary	6.6	5.5
55	The percentage of nouns occurring less than 7 times in the spoken language	15	8

[4] In the chart the computer numeration has been preserved though for different reasons unessential results have been left out.

Table 1. Continuation

56	The percentage of nouns occurring less than 15 times in the spoken language	24	9
57	The percentage of nouns with abstractness 1	24	11
58	The percentage of nouns with abstractness 2	43	12
59	The percentage of nouns with abstractness 3	33	14
60	The percentage of nouns in a paragraph occurring less than 7 times in the 6th grade physics textbook	67	14
61	The percentage of nouns in a paragraph occurring less than 7 times in the 7th grade physics textbook	43	17
62	The percentage of nouns in a paragraph occurring less than 7 times in the 8th grade physics textbook	56	16
63	The percentage of nouns in a paragraph occurring less than 7 times in the 9th grade physics textbook	25	17
64	The percentage of nouns in a paragraph occurring less than 15 times in the 6th grade physics textbook	74	13
65	The percentage of nouns in a paragraph occurring less than 15 times in the 7th grade physics textbook	57	18
66	The percentage of nouns in a paragraph occurring less than 15 times in the 8th grade physics textbook	64	16
67	The percentage of nouns in a paragraph occurring less than 15 times in the 9th grade physics textbook	33	19
77	The average number of letters in words	8.0	0.4
78	The average number of words in sentences	13.1	1.5
79	The average number of letters in sentences	105	11
80	The average number of words in independent sentences	12.7	1.6
81	The average number of letters in independent sentences	102	12
82	The average number of words in the parts of sentences	9.1	0.8
83	The average number of letters in independent sentences	73	8
84	The percentage of the parts of sentences with a length of more than 5 words	76	8
85	The percentage of the parts of sentences with a length of more than 6 words	67	9
86	The percentage of the parts of sentences with a length of more than 7 words	58	9
87	The percentage of the parts of sentences with a length of more than 8 words	49	9
88	The percentage of the parts of sentences with a length of more than 9 words	40	9
89	The percentage of the parts of sentences with a length of more than 10 words	32	7
90	The percentage of the parts of sentences with a length of more than 11 words	26	6
91	The ratio of participial-constructions number to the number of sentences	0.2	0.1
92	The ratio of participial-constructions number to the number of parts of sentences	0.15	0.06

Table 1. Continuation

103	The percentage of infinitives as subjects	4.5	2.9
104	The percentage of predicates with the negative prefix	6.2	4.0
105	The percentage of predicates in the passive	8.6	5.2
112	The number of illustrations per 1000 words	5.5	3.6
113	The number of formulas per 1000 words	50	64
114	The number of symbols per 1000 words	50	54
116	The percentage of scientific-technical nouns in nouns	79	9
117	The average nouns' occurrence in the physics textbooks vocabulary of the 6th to 8th grades	168	97
121	The number of nouns in parts of sentences	3.3	0.6
126	The students' final achievement level (SFA)	54.4	10.7
127	The SFA after equalizing the students' abilities	53.7	10.4
128	The arithmetical mean of the SFA of different variants	53.8	10.3
144	The students' opinion of the paragraph-interest	1.55	0.16
145	The number of various nouns in a paragraph	68	22
146	The recurrence of nouns in a paragraph	2.61	0.5
147	The percentage of the nouns in a paragraph	36	5
153	The paragraph effectivity	0	1.7
207	The number of words used by the computer	462	148
237	The percentage of nouns from the list of 4000 most frequent words	53.4	10.5
238	The percentage of nouns from the list of 3500 most frequent words	86.1	8.2
239	The average occurrence of nouns in the spoken language dictionary by Buchštab	24.4	14.3
240	The average number of selections by Buchštab	17.0	9.3
242	The percentage of the nouns that do not exist in the Zasorina dictionary (4000 most...)	43.2	10.5
243	The percentage of the nouns that do not exist in the Steinfeld dictionary (4000 most...)	54	13
245	The percentage of the nouns that do not exist in the spoken language dictionary	57	10
247	The percentage of the nouns that do not exist in the Buchštab dictionary	20.4	6.3
248	The percentage of the nouns with occurrence less than 10 in the Buchštab dictionary	61.1	9.0
250	The percentage of the noun that do not exist in the scientific-technical dictionary (4000...)	13.9	8.1
252	The percentage of everyday nouns by Sozin	35	11
253	The percentage of the nouns with a statistical value less than 10 in the Zasorina dictionary	57.4	8.9
254	The percentage of the nouns with a statistical value less than 10 in the Josselson dictionary	45.1	10.4
256	The percentage of the nouns with a statistical value less than 10 in the spoken language dictionary	45.1	10.4

Table 1. Continuation

257	The percentage of the nouns with an average statistical value less than 10	43.1	9.5
259	The number of different nouns in the text	67.4	20.8
260	The number of concepts by Sozin	75.5	36.2
261	The percentage of sentences with a length up to 5 words	12.4	5.6
262	The percentage of sentences with a length up to 7 words	22.5	7.3
263	The percentage of sentences with a length up to 9 words	32.9	8.9
264	The percentage of sentences with a length up to 11 words	46.1	10.6
265	The percentage of sentences with a length up to 13 words	57.1	10.4
266	The percentage of sentences with a length up to 15 words	67	9
267	The percentage of sentences with a length up to 17 words	75	8
268	The percentage of sentences with a length up to 19 words	82	7
269	The percentage of sentences with a length up to 24 words	92	6
270	The percentage of sentences with a length up to 29 words	97	4
271	The total number of sentences in a text	36.6	10.8
272	The percentage of the situations where there are no words between two verbs	7.1	5.3
273	The percentage of the situations where there is 1 word between two verbs	2.9	2.6
274	The percentage of the situations where between two verbs there are 2 words	3.2	3.0
275	The percentage of the situations where between two verbs there are 3 words	5.1	3.5
276	The percentage of the situations where between two verbs there are 4 words	6.8	4.0
277	The percentage of the situations where between two verbs there are 5 words	6.9	4.1
278	The percentage of the situations where between two verbs there are 6 words	7.3	3.8
279	The percentage of the situations where between two verbs there are 7 words	6.3	4.3
280	The percentage of the situations where between two verbs there are 8 words	6.5	3.7
281	The percentage of the situations where between two verbs there are 9 words	5.6	3.4
282	The percentage of the situations where between two verbs there are 10 words	5.5	2.9
283	The percentage of the situations where between two verbs there are 11 words	4.7	3.6
284	The percentage of the situations where between two verbs there are 12 words	4.0	2.5
285	The percentage of the situations where between two verbs there are 13 words	4.4	3.8
286	The percentage of the situations where between two verbs there are 14 words	3.3	2.6

Table 1. Continuation

287	The percentage of the situations where between two verbs there are more than 14 words	21.1	9.2
302	The percentage of the nouns that do not exist in the spoken language dictionary	38.7	5.7
303	The percentage of the nouns that exist 1-10 times in the spoken language dictionary	28	6.7
304	The percentage of the nouns that exist 11-30 times in the spoken language dictionary	19.7	5.7
305	The percentage of the nouns that exist 31-80 times in the spoken language dictionary	8.5	3.9
306	The percentage of the nouns that exist more than 80 times in the spoken language dictionary	5.7	3.0
308	The average occurrence of nouns in the spoken language dictionary	21.4	9.0
309	The average occurrence of nouns in the text	2.1	0.4
310	The percentage of the nouns that do not exist in the spoken language dictionary of all words	39	6.5
311	The percentage of the nouns that exist 1-10 times of all words in the spoken language dictionary	59	10
312	The percentage of the nouns that exist 11-30 times of all words in the spoken language dictionary	63	11
313	The percentage of the nouns that exist 31-80 times of all words in the spoken language dictionary	57	12
314	The percentage of the nouns that exist more than 80 times of all words in the spoken language dictionary	8.4	4.4
315	The percentage of the nouns in the text	38.9	3.7
316	The percentage of the verbs that do not exist in the spoken language dictionary	21.1	7.3
317	The percentage of the verbs that exist in the spoken language dictionary 1-10 times	30.5	9.5
318	The percentage of the verbs that exist in the spoken language dictionary 11-30 times	20.7	6.4
319	The percentage of the verbs that exist in the spoken language dictionary 31-80 times	8.1	4.6
320	The percentage of the verbs that exist in the spoken language dictionary more than 80 times	19.0	8.0
321	The total number of verbs	45.2	16.7
322	The average occurrence of the verbs in the spoken language	313	246
323	The average occurrence of the verbs in the text	1.3	0.1
324	The percentage of the verbs that do not exist in the spoken language dictionary of all words	5.4	1.9
325	The percentage of the verbs that exist 1-10 times of all words in the spoken language dictionary	16.1	5.7
326	The percentage of the verbs that exist 11-30 times of all words in the spoken language dictionary	17.8	7.6

Table 1. Continuation

327	The percentage of the verbs that exist 31-80 times of all words in the spoken language dictionary	14.8	81
328	The percentage of the verbs that exist more than 80 times of all words in the spoken language dictionary	7.3	2.8
329	The percentage of the verbs in the text	9.7	1.7
330	The percentage of the adjectives that do not exist in the spoken language dictionary	69	8
331	The percentage of the adjectives that exist 1-10 times in the spoken language dictionary	16.3	7.0
332	The percentage of the adjectives that exist 11-30 times in the spoken language dictionary	7.7	5.5
333	The percentage of the adjectives that exist 31-80 times in the spoken language dictionary	4.1	4.0
334	The percentage of the adjectives that exist more than 80 times in the spoken language dictionary	2.0	2.1
335	Total number of adjectives	58.8	23.7
336	The average occurrence of adjectives in the spoken language	10.7	8.3
337	The average occurrence of adjectives in the text	1.4	0.3
338	The percentage of the adjectives that do not exist in the spoken language dictionary of all words	22.9	5.8
339	The percentage of the adjectives that exist 1-10 times of all words in the spoken language dictionary	11.2	5.0
340	The percentage of the adjectives that exist 11-30 times of all words in the spoken language dictionary	8.1	5.4
341	The percentage of the adjectives that exist 31-80 times of all words in the spoken language dictionary	8.5	6.5
342	The percentage of the adjectives that exist more than 80 times of all words in the spoken language dictionary	1.1	1.0
343	The percentage of the adjectives in the text	12.7	2.9
344	The percentage of the adverbs that do not exist in the spoken language dictionary	17.3	8.4
345	The percentage of the adverbs that exist in the spoken language dictionary 1-10 times	13.6	9.3
346	The percentage of the adverbs that exist in the spoken language dictionary 11-30 times	9.7	7.1
347	The percentage of the adverbs that exist in the spoken language dictionary 31-80 times	7.1	6.3
348	The percentage of the adverbs that exist in the spoken language dictionary more than 80 times	52.7	12.2
350	The average occurrence of adverbs in the spoken language dictionary	501	232
351	The average occurrence of adverbs in the text	1.3	0.2
352	The percentage of the adverbs that do not exist in the spoken language dictionary of all words	2.4	1.4
353	The percentage of the adverbs that exist 1-10 times of all words in the spoken language dictionary	3.9	3.4

Table 1. Continuation

354	The percentage of the adverbs that exist 11-30 times of all words in the spoken language dictionary	4.8	3.9
355	The percentage of the adverbs that exist 31-80 times of all words in the spoken language dictionary	7.1	6.9
356	The percentage of the adverbs that exist more than 80 times of all words in the spoken language dictionary	12.5	13.7
357	The percentage of the adverbs in the text	5.3	1.8
358	The percentage of the prepositions that do not exist in the spoken language dictionary	0.02	0.16
359	The percentage of the prepositions that exist 1-10 times in the spoken language dictionary	5.4	5.2
360	The percentage of the prepositions that exist 11-30 times in the spoken language dictionary	1.8	2.2
361	The percentage of the prepositions that exist 31-80 times in the spoken language dictionary	3.9	3.6
362	The percentage of the prepositions that exist more than 80 times in the spoken language dictionary	89	7
363	The total number of prepositions in the text	44.2	13.7
364	The average occurrence of prepositions in the spoken language dictionary	2333	392
365	The average occurrence of prepositions in the text	3.1	0.6
367	The percentage of prepositions that exist 1-10 times of all words in the spoken language dictionary	3.1	3.4
368	The percentage of prepositions that exist 11-30 times of all words in the spoken language dictionary	1.5	1.9
369	The percentage of prepositions that exist 31-80 times of all words in the spoken language dictionary	7.3	7.5
370	The percentage of prepositions that exist more than 80 times of all words in the spoken language dictionary	33.2	5.0
371	The percentage of the prepositions in the text[5]	9.7	1.3
400	The percentage of the words in the text that do not exist in the spoken language dictionary	38.7	3.9
401	The percentage of the words in the text that exist in the spoken language dictionary 1-10 times	18.2	3.2
402	The percentage of the words in the text that exist in the spoken language dictionary 11-30 times	11.7	2.4
403	The percentage of the words in the text that exist in the spoken language dictionary 31-80 times	5.6	1.9
404	The percentage of the words in the text that exist in the spoken language dictionary more than 80 times	25.4	4.8
406	The average occurrence of words of the text in the spoken language dictionary	692	100

[5] Similar characteristics to those of 358-371 of the pronouns and conjunctions were not much in correlation with the indexes showing the effectivity of the text. That is why their list is not given here.

Table 1. Continuation

407	The average occurrence of words in the text	1.8	0.2
424	The length of sentences in words	13.0	1.5
425	The length of sentences in letters	105	11
426	The number of numerals, symbols formulas and abbreviations in 100 words	6.4	3.5
427	The number of rubrications in 100 words	0.04	0.15
428	The number of numerals in 100 words	1.5	0.9
429	The number of abbreviations in 100 words	1.0	0.5
430	The number of symbols in 100 words	1.5	0.8
431	The number of formulas in 100 words	0.9	1.2
432	The number of foreign words in 100 words	0.1	0.2
433	The number of abbreviations in 100 words	1.2	1.3
435	Effectivity modification	14.3	9.1
436	The percentage of concepts according to Sozin	16.2	4.7
437	The number of noun constructions with a length of 1 word in 100 words	17.8	2.8
438	The number of the noun constructions with a length of 2 words in 100 words	6.2	1.2
439	The number of the noun constructions with a length of 3 words in 100 words	2.0	0.9
440	The number of the noun constructions with a length of 4 word in 100 words	0.6	0.4
441	The number of the noun constructions with a length of 5 word in 100 words	0.2	0.2
448	The percentage of the scientific-technical concepts out of all words	31	4.7

cept 252). The third method was completely automated. All the texts were typed in the computer and analyzed by programs of morphological analysis that had been worked out in Kiev by N.P. Darčuk and her colleagues (in Verbickij 1984). Then the computer determined the frequency of every word in our paragraphs relying on the computer frequency list of the Russian speech[6]. The computer also counted the words in the paragraphs and the letters in the words. So were determined the values of characteristics[7] No 207-236 and 261- 447 (except 436).

Text Effectivity Factors of Physics Textbooks

In this experimental research we have fixed three students' final achievement indexes (indexes 126, 127, 128 in table 1), the index of paragraph interest (index

[6] The frequency list of words of the Russian speech was elaborated in Moscow University by Buchštab and colleagues.

[7] T. Tamman and L. Urm participated in this work.

144) and the integral index of the paragraph effectivity (index 153) that was calculated by formula (1). All the three factors of student final achievement had linear correlations between themselves and they were equal to 0.99. It means that these indexes are identical and we may use one of them. Thus the average level of students' abilities in small groups is equal to the total average level. The unequal distribution of the variants has not brought about any noticeable deviation of the results.

We calculated linear correlations between all the text characteristics and learning effectivity indexes for finding the learning resultativity dependence on the text characteristics. The text characteristics influence the independent text-learning process if the correlation index is reliable. The reliable correlation coefficients have been given in table 2.

Let us have a look at what kind of paragraph characteristics the *students` final achievement* level depends on. The correlation coefficient was relatively great between the students' final achievement level and the percentage of the nouns in a paragraph (index 147). The greater percentage of nouns in the text conduces its learning because it minimizes the percentage of adverbs and participles in the text. But these two types of words make the text more difficult to understand (Granowsky, Botel 1974). Clear texts have comparatively many verbs and nouns (Wiio 1968).

The second correlation coefficient in absolute value with student final achievement level was found at the average abstractness of nouns in paragraph (index 40). The greater is the abstractness of the text the lower is the level of student final achievement.

How the abstractness of the text influences its understanding was studied by many researchers some decades ago already (Flesch 1950; Mikk 1974).

The level of abstractness of the nouns was measured in our research by a 3-mark scale:

1 - nouns that mark the things and living creatures that a man can directly accept with sensory organs (for example: car, child);

2 - nouns that mark the phenomena that a man can directly accept with sensory organs (for example: light, sound);

3 - nouns that mark the mental constructions that a man cannot directly accept with sensory organs (for example: function, subject). The nouns with the mark 1 were called concrete and with the mark 3 - abstract.

In table 2 we can see that the percentage of the nouns essentially influences the students' final achievements (SFA). The greater is the percentage of concrete nouns the better is the SFA. The greater is the percentage of abstract nouns the smaller is the SFA. Abstract words make the text more difficult to understand because they have little connection with the students' everyday life.

The level of abstractness of the nouns, the percentage of the nouns and verbs in a paragraph are the most important text factors that influence the SFA. In table 2 there is a relatively small number of factors that influence the SFA. But there are more factors that are in quite a high correlation with the students' opinion of the paragraph interest.

The *students' interest* in physics texts depends on many characteristics regarded, first of all, on the vocabulary. In its absolute value the correlation was highest between the rate of the interesting text and the words frequency of occurrence in a paragraph (index 407). The more often words occur in the text the less students want to study it. The next correlations in quantity are close to this in their essence. The more often words typical of the physics textbooks occur in the paragraph given the less is the students` interest in it (indexes 39, 44). This effect is not common. The higher noun frequency of occurrence makes the text more understandable and that must increase the students' interest in the text. It is possible that the same nouns are used too often in the physics textbooks (the average value - 400 times in five steps). Similar vocabulary makes examples and supplements for students not very interesting. It is also possible that some students do not like the physics terms.

That tendency was characteristic to other factors too (indexes 62, 63, 65, 66, 67). The last indexes mean that if in the text there are more nouns with a small frequency of occurrence in physics textbooks it is more interesting for the students.

The next factors (indices 46, 116, 436, 448): scientific-technical terms - decrease the students interest in the text if they are used too often in it. The frequency of occurrence of scientific-technical words was fixed by the adequate dictionary (Denisov, Morkovkin, Safjan 1978).

This research has confirmed the hypothesis that the quantity of symbols in the text influences the students' interest in the text. Quantities or concepts marked with letters were considered symbols. We have found out that a lot of symbols and formulas in the text make it uninteresting for the students (indices 32, 33, 113, 114, 430, 431).

The text interest is considerably influenced by the average abstractness of the nouns (index 40). This influence is caused by proportion of the abstract nouns:

its high value decreases the text interest of the students (index 59).

Table 2
The effectivity factors of the physics textbook

SFA - students' final achievement level (index 126)
TI - text interest (index 144); TE - text effectivity (index 153)

No	Name of the factor	Correlation coefficient with		
		SFA	TI	TE
32	The number of formulas in a paragraph		-0.53	-0.38
33	The number of symbols in a paragraph		-0.57	-0.48
39	The average paragraph noun occurrence in the physics textbook		-0.70	-0.60
40	The average abstractness of nouns in a paragraph	-0.53	-0.52	-0.61
41	The average paragraph noun occurrence in the physics textbook for grade 6.		-0.44	
42	The average paragraph noun occurrence in the physics textbook for grade 7.		-0.58	-0.41
43	The average paragraph noun occurrence in the physics textbook for grade 8.		-0.49	-0.44
44	The average paragraph noun occurrence in the physics textbook for grade 9.		-0.69	-0.59
46	The average scientific-technical noun occurrence in paragraph		-0.57	-0.51
50	The percentage of 12-and-more-letter nouns	0.40		
52	The percentage of 14-and-more-letter nouns		0.36	0.43
57	The percentage of concrete nouns	0.51		0.47
59	The percentage of abstract nouns	-0.40	-0.48	-0.51
61	The percentage of nouns in a paragraph occurring less than 7 times in the 8th grade physics textbook		0.62	0.50
62	The percentage of nouns in a paragraph occurring less than 7 times in the 8th grade physics textbook		0.38	0.39
63	The percentage of nouns in a paragraph occurring less than 7 times in the 9th grade physics textbook		0.67	0.53
65	The percentage of nouns in a paragraph occurring less than 15 times in the 7th grade physics textbook		0.60	0.46
66	The percentage of nouns in a paragraph occurring less than 15 times in the 8th grade physics textbook		0.40	0.40

Table 2. Continuation

67	The percentage of nouns in a paragraph occurring less than 15 times in the 9th grade physics textbook		0.65	0.52
79	The average number of letters in sentences		-0.40	
113	The quantity of formulas in 1000 words		-0.46	
114	The number of symbols in 1000 words		-0.50	-0.40
116	The percentage of scientific-technical nouns		-0.64	-0.49
117	The average nouns' occurrence in the physics textbooks vocabulary of the 6th to 8th grades		-0.59	-0.49
121	The number of noun parts of sentences	0.52		0.48
147	The percentage of the nouns in a paragraph	0.54		0.51
237	The percentage of the nouns from the list of 4000 most frequent words		-0.55	-0.38
238	The percentage of the nouns from the list of 3500 most frequent		-0.44	-0.39
239	The average occurrence of nouns in the spoken language dictionary by Buchštab		-0.37	-0.41
240	The average number of selections by Buchštab	-0.34	-0.37	-0.44
250	The percentage of the nouns that do not exist in the scientific-technical dictionary		0.45	0.41
252	The percentage of everyday nouns	0.32	0.68	0.58
260	The number of concepts by Sozin		-0.39	-0.34
264	The percentage of sentences with a length of up to 11 words		0.40	
265	The percentage of sentences with a length of up to 13 words		0.40	
266	The percentage of sentences with a length of up to 15 words		0.40	
267	The percentage of sentences with a length of up to 17 words		0.45	
268	The percentage of sentences with a length of up to 19 words		0.37	
273	The percentage of the situations where there is 1 word between two verbs		0.36	0.43
287	The percentage of the situations where between two verbs there are more than 14 words		-0.41	
302	The percentage of the nouns that are not in the spoken language dictionary		0.48	0.49
306	The percentage of the nouns that exist more than 80 times in the spoken language dictionary		-0.35	
310	The percentage of the nouns that do not exist in the spoken language dictionary of all words	0.45	0.47	0.56
312	The percentage of the nouns that exist 11-30 times of all words in the spoken language dictionary		-0.55	-0.39
314	The percentage of the nouns that exist more than 80 times in the spoken language dictionary		-0.35	

Table 2. Continuation

317	The percentage of the verbs that exist in the spoken language dictionary 1-10 times		-0.36	
318	The percentage of the verbs that exist in the spoken language dictionary 11-30 times		0.35	
326	The percentage of the verbs that exist 11-30 times of all words in the spoken language dictionary		0.34	
329	The percentage of the verbs in the text		0.35	
332	The percentage of the adjectives that exist 11-30 times in the spoken language dictionary		0.46	
337	The average occurrence of adjectives in the text		-0.43	-0.49
340	The percentage of the adjectives that exist 11-30 times of all words in the spoken language dictionary		0.38	
344	The percentage of the adverbs that do not exist in the spoken language dictionary	-0.33	-0.34	-0.37
346	The percentage of the adverbs that exist in the spoken language dictionary 11-30 times		0.45	0.33
354	The percentage of the adverbs that exist 11-30 times of all words in the spoken language dictionary		0.47	0.37
360	The percentage of the prepositions that exist in the spoken language dictionary 11-30 times		-0.37	
362	The percentage of the prepositions that exist in the spoken language dictionary more than 80 times		0.33	
368	The percentage of the prepositions that exist 11-30 times of all words in the spoken language dictionary		-0.35	
371	The percentage of the prepositions in the text	-0.33		-0.32
407	The average occurrence of words in the text		-0.72	-0.68
426	The number of numerals, symbols, formulas and abbreviations in 100 words		-0.50	-0.44
430	The number of symbols in 100 words		-0.52	-0.53
431	The number of formulas in 100 words		-0.54	-0.43
436	The percentage of concepts according to Sozin		-0.58	-0.49
448	The percentage of the scientific-technical concepts out of all words		-0.57	-0.37

The students' interest in the text depends more on factors than their final achievement. Some factors make the text interesting but have no influence on how comprehensible it is or how easily it is mastered. But if a factor raises the students' final achievement as a rule it will also raise the students` interest in the text. The correlation of factors with SFA and TI have the same sign in table 2.

It is surprising how the percentage of the nouns that occur in speech influences the effectiveness of the text: the greater is the number of the spoken language nouns in the physics texts the lower is the students' interest in these texts (characters 302, 310 and also 237). Thus the result puts in doubt the recommendation given by one of the dictionaries authors to rely on the 4000 most frequent words while writing schoolbooks (4000..., 1986). If we prefer words out of the most frequent 4000 words the tenth-form students' interest in the text will be higher. Obviously 4000 words are not enough to evaluate tenth-form students' schoolbooks - their vocabulary is much larger. The indices of the correlation of every-day nouns were traditional only with expert I. Sozin. He marked everyday nouns in paragraphs and their higher percentage raised paragraph effectivity (character 252).

The *text effectivity* depends on the same factors as the SFA and the students' interest of text. Therefore we don`t analyze these factors.

We have two ways to use the factors of effectivity of learning texts:

1) By its values we can estimate the quality of learning a text. It is also important to pay attention to the factors (Table 1) that had no reliable correlations with the indices showing the effectivity of the text, e.g., the frequency of participles in the text (indices 91, 92), the frequency of the passive voice and the negative sentences (indices 104, 105), the frequency of long substantive constructions (indices 440, 441) had no influence on the effectivity of the text, though psycho-linguistical studies have shown that such kind of constructions may make it more difficult to understand sentences. The text as a whole does not show the effect of all the possible factors of difficulty. One of the reasons for this may be that these factors occur in the text accidentally and rarely. It confirms the statement that the rules of a good text must be more extensive than the recommendations we can get in Table 2.

2) By using these factors it is possible to give recommendations for raising the textbook effectivity.

In order to estimate physics texts we have worked out some formulas in traditional way by regression analysis.

In the process of elaborating formulas we considered not only statistical validity of arguments but also their content.

The formula for prognosticating the summary effectivity of the physics texts worked out by us is the following:

(2) $y_{153} = -3.17X_{40} - 0.0103X_{43} + 0.0582X_{310} - 1.81X_{337} - 0.320X_{371} + 10.68$

In this formula

X_{40} - the average abstractness of nouns in a paragraph;
X_{43} - the average paragraph noun occurrence in the physics textbook for grade 8;
X_{310} - the percentage of the nouns that do not exist in the spoken language dictionary of all words;
X_{337} - the average occurrence of adjectives in the text;
X_{371} - the percentage of the prepositions in the text.

In this formula X_{310} has not a common sign. The available reasons of this effect are described some pages before.

The multiple correlation coefficient of formula (2) is equal to R = 0.88. It is a good result: this formula prognosticates about 80 per cent of the average effectivity of independent work with the physics textbooks in the senior forms.

The SFA level is predicated by the next formula:

(3) $y_{126} = 60.6 - 16.9X_{40} + 0.776X_{147}$

In this formula

X_{40} - the average abstractness of nouns in a paragraph;
X_{147} - the percentage of the nouns in a paragraph.

The multiple correlation coefficient of formula (3) is equal to R = 0.64.

The next prognosticating formula has been worked out for the text interest:

(4) $y_{144} = - 0.0034X_{79} - 0.020X_{360} - 0.45X_{407} - 0.042X_{431} - 0.0077X_{448} + 2.95$

In this formula

X_{79} - the average number of letters in sentences;
X_{360} - the percentage of the prepositions that exist in the spoken language dictionary 11 - 30 times;
X_{407} - the average occurrence of words in the text;
X_{431} - the number of formulas in 100 words;
X_{448} - the percentage of the scientific-technical concepts out of all words.

The multiple correlation coefficient of the formula is equal to R = 0.86. It means that formula (4) prognosticates the student interest in the text in the limits of 74 per cent. In reality the exactness of the prognosis would be smaller because in

this research some important factors of learning texts were not taken into consideration (for example the subject matter).

We can use formulas (2)-(4) all together or separately. We need not use the whole text while analyzing it with the help of the formulas described above. The previous researches have predicted that for quite an exact prognostication it is important to use about 30 paragraphs to fix the factors for these formulas.

Conclusion

This article is dedicated to working out the methods of analyze effectivity physics textbook. In the experimental part of the study we tried to get the more exact opinion of the difficulty of the basic texts. For this purpose eighty questions on each section were made up and given to different students in eight variants. Alongside with the right answers to the questions we were interested in the students` different opinion of the text. The basic texts were longer than usual - 500 words on the average.

While analyzing the text most of the work was done by computers using the morphological analysis program of the Russian language. It gave us different characteristics of the occurrence of words .

The results of the study were unexpected because they showed that the more often the text had the words used in a physics textbook the lower was its effectivity. The main reason for the fall of the interest level was the high occurrence of the words in the section.

Evidently the vocabulary of the Russian physics textbooks is too limited and the amount of special terminology too great.

The abstractness of the nouns formed the next group of factors showing the effectivity of the physics textbooks. The result has been confirmed repeatedly in earlier studies.

The final achievement was too little influenced by the length of a sentence and it was also unexpected.

Relying on the factors of effectivity found we worked out the formulas for prognosticating the effectivity of a physics text. Three of those have been given in the article.

The prognosticating exactness of these formulas is high, the multiple correlation coefficient is about 0.86. Despite that we have to take into consideration that the real effectivity of a textbook is somewhat different from the effectivity we get while prognosticating the textbook.

This research has not taken into consideration all the factors of the paragraph effectivity in physics. For example we did not investigate the logical structure, methodical system, illustrations, the system of tasks of a textbook, etc. The textbook was investigated in the course of the students' independent learning. But teachers' explanations may change the influence of the textbook characteristics on the effectivity of the learning process.

Despite these remarks we hope that the formulas enable us to get the first assessment of schoolbook effectivity before we begin to experiment the textbook at school.

References

Bamberger R., Rabin A. T. (1984). New Approaches to Readability: Austrian Research (LIX readability index). *Reading Teacher 37, 512-519.*

Buchovcev B.B., Klimontovič J.L., Mjakišev, G.J. (1984). *Fizika. Učebnik dlja 9 klassa srednej školy* (Physics. Textbook for Form 9 of Secondary School). Moscow: Prosveščenie .

Denisov P.N., Morkovkin V.V., Safjan J.A. (1978). *Compound Frequency Dictionary of 3047 Words of Russian Scientific and Technical Vocabulary.* Moscow: Russkij jazyk (in Russian).

Flesch R.F. (1950). Measuring the Level of Abstraction. *Journal of Applied Psychology 34, 384-390.*

Glass G.V., Stanley J.C. (1970). *Statistical Methods in Education and Psychology.* Englewood Cliffs, N.J., Prentice Hall.

Granowsky A., Botel M. (1974). Background for a Syntactic Complexity Formula. *Reading Teacher 28, 31-35.*

Klare G.R. (1963). *The Measurement of Readability.* Iowa: Iowa State University.

Klare G.R. (1974-75). Assessing Readability. *Reading Research Quarterly 10, 62-102.*

Mackovski M.S. (1976). Problemy čitabelnosti pečatnogo materiala (Problems of Readability of the Printed Material). In: Dridze, T.M., Leontev, A.A.(eds.), *Substantial perception of spoken information.* Moscow, Nauka: 126-142.

Mikk J.A. (1974). Metodika razrabotki formul čitabelnosti (Methods of Elaborating Readability Formulas). *Soviet Pedagogy and School (Tartu) 9, 78-163.*

Mjakišev G.J., Buchovcev B.B. (1985). *Fizika. Učebnik dlja 10 klassa srednej školy* (Physics. Textbook for Form 10 of Secondary School). Moscow, Prosveščenie.

Průcha J. (1986). K razrabotke parametrov složnosti učebnogo teksta (On treating the parameters of the complexity of study text. *Problemy školnogo učebnika (Moscow) 15, 143-164.*

Schuyler M.R. (1982). Readability Formula Program for Use on Microcomputers. *Journal of Reading 25, 560-591.*

Verbickij, V.A. (ed.) (1984). *Avtomatizacija analiza naučnogo teksta* (Automatization of the analysis of scientific texts). Kiev, Naukova dumka.

Wiio O.A. (1968). *Readability, Comprehension and Readership.* Tampere, Acta Universitatis Tamperensis A, 22.

4000 najbolee upotrebitel'nych slov russkogo jazyka (4000 most often used words in Russian) 1986:5. Moscow, Russkij Jazyk.

Glottometrika 14, 1993, 104-112

The Dependence of the Learning Time on the Text Characteristics

Hasso Kukemelk, Tartu

Learning time is an important variable of a learning process. It shows how effectively a student can acquire a lesson.

In our epoch rich in information it is very important to study a material in the shortest possible time. Human life is limited and we cannot use up too much time on students mastery learning.

According to J.B. Carroll learning time depends on a student's abilities, the quality of instructions and instructional quality (Carroll 1963: 729-730). Instructional quality, in its turn, depends greatly on the textbook's characteristics. This article studies the quality of a text more thoroughly.

The problems of a text have been studied by several Russian and Estonian researchers. J. Mikk has studied how complicated the text presentation is and how it influences the learning effectivity (cf. Kukemelk & Mikk 1993). A.M. Sochor has studied the relation of the logical structure of a text with the students' achievement (Sochor 1974).

The following part of the article estimates how the text learning time depends on its characteristics. We solved the problem experimentally. Students learned some texts, we measured the learning time and then investigated which text characteristics influenced the time.

The course of the experiment

An experiment was carried out with 102 17-years-old the eleventh form secondary school students. The students for the experimental research were chosen by the table of random numbers. There were 54 girls and 48 boys in the sample. The students were to learn 40 mathematics, physics and astronomy texts independently. Every text had about 2600 (the standard deviation S = 236) letters. Each time at the beginning of a text learning the students knowledge was fixed.

At the end of the process the students' achievement was fixed. The final achievement was fixed with the test the structure of which was similar to the test for checking preliminary knowledge. The time used for learning was fixed quite exactly. The mistake was up to a minute. Every student was to learn for such a time that he got 80 per cent of points at the end of the learning process. The students' final achievement level was taken equal to 80 per cent of points because some researches (Block, Anderson 1975; Cristoffersson 1971:135) have found that about 80 per cent is an optimal level of acquiring of learning text. If the student's result was lower the text was returned to him and he had to study it again. So it could happen three times. Every time the learning time was fixed and added to the total time. If a student could not achieve the necessary points during three periods the experimenter asked him orally and told him that he had achieved the aim and could start with the next text. But the experiment did not take such a learning time into consideration because its mistake was too big and the student had not achieved the demanded level of knowledge independently.

The students' abilities were fixed with the Estonian variant of Amthauer IQ test - AS-test. That test had 9 subtests and which enable to fix various students' abilities. In our experiment the arithmetical mean of IQ-s was 103. As the normal mean for IQ being 100 we may admit that the students in that experiment were of average abilities (the standard deviation of a students abilities was 9).

The correlation between the learning time and the text characteristics

Alongside with experimental fixing of the learning time we found out several characteristics that would influence it. The dependence of the text learning time on the text characteristics was found by means of correlation analysis the results of which are presented in table 1. All correlation coefficients in that table are statistically significant (at the 95 percentage level).

All these mentioned characteristics except the rate of the interesting text increase the learning time. Only by making the text more interesting helps us to decrease the time necessary for learning. Here there are some possible reasons:

1) It is quite difficult to compile a complicated text in an interesting way (the text would not be interesting if a student could not understand it).

2) An interesting lesson fascinates a student`s attention. He is more deeply engaged in the text and thus he can achieve good results in a shorter learning

3) The laws and facts in an interesting text are so different from the rest of the information that the students master it more quickly (these parts of sentences irritate the human brain more than the usual information).

These rules may be combined in a real learning process at school.

The rate of the interesting text (the characteristic 12) in correlation with the learning time was the *highest.* Partly it could have been due to the process of carrying out the experiment (students were large to master the material very quickly). Therefore we should consider the fact that an interesting text is equal to the text quickly mastered.

The next correlation coefficients of the learning time are with the text graph characteristics 7 and 8. It says that learning time depends essentially on the concept number and their logical structure (8). The logical graph shows how the concepts of the paragraph are connected with each other. So the graph nodes are connected concepts and the graph edges are the logical connections between them. If there are a lot of concepts in the text and they have a lot of relations between themselves, then a student has to learn that text for a long time. In that way the text with a complicated structure takes a lot of time to be mastered.

The learning time also depends on the percentage of nouns (10) in the text because a lot of nouns in the sentence make its structure more complicated and it takes much time to learn. The learning time depends in the same way on the number of punctuation marks (3). A lot of punctuation marks refer to dependent clauses, a list of various things and so on in the sentence. So the number of punctuation marks shows how complicated the sentence structure is.

Here is performed an example of physical learning text and its logical graph (see Fig. 1)

Rest energy in classical mechanics is an entirely unexpected quantity. This is the energy that is stored into particles at the moment they are formed. It is their internal characteristic. In mechanical, electrical and chemical processes this energy is unnoticed as the forces taking part neither create nor destroy elementary particles of the substance and the rest energy remains unchanged. In nuclear processes only a small part of rest energy becomes transformed into other forms of energy. In transforming processes of elementary particles the rest mass can disappear altogether. In the process of annihilation of particles and antiparticles their rest mass is transformed into the energy of photons with zero rest energy.

Table 1
The dependence of the learning time on the text characteristics

No	Characteristics	R	M	S
1	The percentage of 10-and-more-letter nouns of the total number of words	0.38	41.3	11.8
2	The percentage of 12-and-more-letter nouns of the total number of words	0.36	25.3	10.2
3	The number of punctuation marks	0.51	93	42
4	The word length in letters	0.36	6.9	0.4
5	The number of the defined concepts and rules	0.45	3.1	2.5
6	The graph rank of the logical structure of the text	0.39	2.6	0.5
7	The number of the graph nodes of a text logical structure	0.64	14.3	4.6
8	The number of the graph edges of a text logical structure	0.66	18.9	8.0
9	The percentage of nouns of the total number of words	0.55	35.6	5.4
10	The percentage of the nouns with an abstractness 2 of the total number of the words	0.37	11.2	8.2
11	The number of the formulas in a 100 words of the text	0.45	0.85	2.11
12	The rate of the interesting text as estimated by the students	-0.84	1.59	0.19

R - the correlation coefficient.
M - the arithmetical mean of the characteristic value.
S - the standard deviation.

Defined concepts and rules in the text are usually to be learned by heart. For that reason their number (5) has quite a high correlation coefficient with the learning time. Formulas have always been a great problem for the students. A formula is a succinct sentence full of information. Every formula is a coded sentence. Before using the formula students have to decode it into an ordinary sentence. For that the student's brain has to do some operations, in addition. Accordingly a lot of formulas in the text make it more difficult to comprehend and the learning process takes more time.

The learning process is longer if there are longer nouns in the sentences. A

longer word contains more information than a short one (Mikk 1981: 35) and it takes more time to learn.

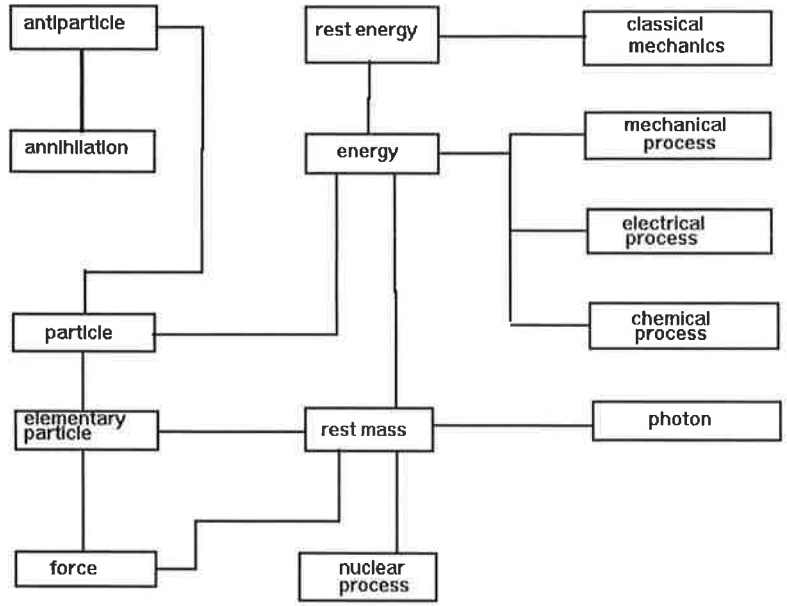


Figure 1. Logical graph of physical learning text

There is quite an essential dependence of learning time on nouns with different abstractness. It can be explained in the following way: nouns with an abstractness 1 are similar to understand and so they do not influence the learning time at all. Nouns with an abstractness 3 are too general and the students cannot understand their full meaning. They study those nouns by heart. That is why nouns with an abstractness 2 influence the learning time.

Using the results in table 1 it is possible to make some conclusions for composing new learning texts in order to avoid the student excessive learning time:

1) the texts must be interesting;

2) use only the terms and rules that cannot be avoided while estimating and mastering;

3) use only necessary connections between the terms and the rules;

4) do not use long sentences especially with some dependent clauses;

5) do not use formulas for all dependencies;

6) use fewer abstract words (especially if you can use concrete ones instead);

7) use shorter words if possible.

The formulas for prognosticating learning time

Regression analysis was used for all text factors (s. table 1). In that way we worked out some formulas in order to prognosticate the learning time for a 1.5-page text. These formulas can be used for prognosticating the learning time of the students whose IQ is between 92 and 112 points. In this interval these formulas give good results in the sciences (physics, mathematics, astronomy).

For practical experience the following formula gives good results:

(1) $t = 0.27X_4 + 0.17X_5 + 0.18X_6 + 0.11X_7 + 2.87$

where t - the learning time in minutes;
X_4 - the number of the formulas in 100 words of the text;
X_5 - the number of the logical graph edges;
X_6 - the percentage of the nouns of all the words;
X_7 - the percentage of the nouns with an abstractness 2 of all the words.

The multiple correlation coefficient for formula (1) is R = 0.79. The other factors in table 1 were not used in formula (1) because they have statistically significant correlations with these factors.

It is possible to analyze the learning process by the boys and the girls separately. In the present situation we have found small difference between the formulas of the boys' and girls' learning time. The ability test fixed some differences between them. It is logical that the boys' and girls' abilities are not identical (the structure of their abilities is different). Thus we got two formulas. The first is the formula of the girls' learning time:

(2) $t = 0.26X_4 + 0.21X_5 + 0.23X_6 + 3.03$

where the multiple correlation coefficient is R = 0.76 and the names of the factors are the same as in formula (1).

The formula of the boys' learning time is the following:

(3) $t = 0.27X_4 + 0.17X_5 + 0.18X_6 + 0.11X_7 + 2.85$

where the multiple correlation coefficient is R = 0.79. As we can see the differences between the formulas (2) and (3) are not very big. The most important difference lies in factor X_7. That factor is not very important for girls.

The experimental research enabled us to make up three formulas related the students' text learning time to its factors.

The most valid formula is (1) for evaluating the textbooks in the science subjects. This conclusion is based on the following assertions:

1) All indices of formula (1) are quite simple and they can be fixed in one and a half pages of a text (about 2600 letters) operatively.

2) The multiple correlation coefficient is quite high (R = 0.79) so the formula describes about 60 percentage of the variance of learning time.

3) Formula (1) has been drawn up on the basis of the average abilities of 11th-form students. It means that this formula enables to analyze the textbooks on the sciences.

4) When we enter the textbook materials into a computer we can use some programs for analyzing the textbook quality and for prognosticating the students' learning time.

This formula also has some disadvantages.

1. It does not take into account the fact that the students differ in their abilities, i.e. we cannot prognosticate the learning time of very slow and very quick learners.

2. The field of the usage of formula (1) is not very large. It is possible to get good results in Form 11 and quite normal results in Form 10 and Form 12 but not in Forms 7-9 because the formula has been drawn up for the students of certain abilities.

3. The formula does not hold the students' abilities. Formula (1) cannot be applied to students with different abilities (bright learners, average learners and less bright learners, especially). The reason is very simple - we haven't found any essential differences between their learning times. If the average learning time of a group was for example 19 minutes, the less bright learners' group spent 21-22 minutes and the bright learners' group 16-17 minutes. But the differences here were quite small so that the formulas for the groups did not have any essential differences.

The prognostication of the learning time in practice

In order to prognosticate the students' learning time in practice we have to enter the numerical values of the factors into the formula.

For example if we want to know how much time boys would spend on learning a two-pages text we have to take formula (3) and find the following factors in a 2600-letter text:

X_4 - the number of formulas in 100 words of the text (in our example it is 1,8);

X_5 - the number of logical graph edges of a text logical structure (27) (this factor we have to find as normed value on 2600-letter text);

X_6 - the percentage of nouns of the total number of words (45);

X_7 - the percentage of the nouns with an abstractness 2 of the total number of the words (24);

Now we can enter these values into formula (3):

$t = 0.27(1.8) + 0.17(27) + 0.18(45) + 0.11(24) + 2.85 = 19$ minutes.

As the length of the text is two-pages we will have to multiply this time by 1.33 (2 divided by 1.5). As a result we prognosticate the real learning time as 25 (19 multiply to 1.33) minutes for boys in case if they have the positive learning motivation.

It is also possible to use the other formulas for prognosticating the learning time in the same way. It is important to fix the numerical values of factors in a

2600-letter text and then, by using these formulas, we can predict with 60-70 per cent probability the learning time needed.

On the basis of the textbook these formulas enable us to predict how many lessons for example, in physics, the students of the certain age group need to master the material independently. Thus we can predict whether the textbook is really suitable for the students or not.

References

Block J.H., Anderson L.W. (1975). *Mastery Learning in Classroom Instruction.* New York, Macmillan.

Carroll J.B. (1963). A Model of School Learning. *Teachers College Record 64, 723-733.*

Cristoffersson N.-O. (1971). *The Economics of Time in Learning.* Malmö.

Kukemelk H., Mikk J. (1993). The prognosticating effectivity of learning a text in physics. *Glottometrika 14, 82-103.*

Mikk J.A. (1981). *Optimizacija složnosti učebnogo teksta.* (Optimizing the complexity of learning text). Moscow: Prosveščenie.

Sochor A.M. (1974). *Logičeskaja struktura učebnogo materiala* (The logical structure of the study material). Moscow: Pedagogika.

Glottometrika 14, 1993, 113-121

Mathematical Verbal Problems: Differences in Solving Difficulties

Madis Lepik, Tallinn

Introduction

A central theme of mathematical instruction is to help students to develop their problem solving skills. The role of a teacher and a designer of instructional materials is to create a learning environment in a way that optimizes the student learning process. According to K. Pfeiffer et al. (1987), the problems of too easy as well as the problems of too difficult level do not inspire students to solve problems. Problems that are too easy are not rewarding to solve, they are just dull routine work. Problems that are too difficult do not guarantee enough success to be inspiring, and students tend to give up. Research indicates that the best results are achieved when success is likely in about 50%.

To find better ways of teaching problem solving skills one has to study the relative difficulty of problems. There is a number of studies that deal with the relationship between the problem variables and the difficulty level in mathematical verbal problems. It was assumed that the relative difficulty of problems could be determined by analyzing the construction of the wording of the problem. Most of such variables (that aid or interfere with the pupil's performance in solving verbal problems) fall into the following groups:

(1) variables describing the textual presentation of the problem - readability variables (Jerman, Mirman 1974; Austin, Lee 1982; Moyer 1984);

(2) variables describing data, calculable values and mathematical operations in the problem - computational variables (Jerman, Mirman 1974);

(3) variables describing the logical structure of the problem - logical variables (Sochor 1974; Lepik 1990);

(4) variables describing the presence or absence of verbal cues that help in selecting operations (Byers, Erlwanger 1985; Nesher, Teubal 1975);

(5) variables describing the interest and problem solving experience of the student (Wright, Wright 1986).

In my previous paper (Lepik 1990) the influence of 31 linguistic, computational and logical variables on algebraic verbal problem solving performance was studied. To examine the logical structure of the problems in greater detail problem graphs were used. This approach enabled to define a number of problem variables used to describe the complexity of graphs in graph theory. To study the influence of problem variables on the proportion of correct strategies and the rate of solving, correlation analysis was used. Six variables were found to be statistically significant in predicting the values of these performance variables. They were structural variables, mostly definable on the basis of the problem graph.

In the present paper we are going to study the influence of the subject's abilities on the significance of problem difficulty variables. According to O.Magne error patterns of high-achievers and low-achievers differ greatly (Magne 1989). V.A. Krutetskii has observed that mathematical achievement may depend on psychological features which are highly individual (Krutetskii 1976). In his contrastive comparisons he observed typical differences between the high-achiever and the low-achiever in three basic aspects:

(1) information gathering, including initial orientation to a problem;

(2) information processing in problem solving;

(3) information retention.

So, the question arises whether the role of the problem variables in building up problem difficulty is the same for students of different achievement levels.

The purpose of the present paper is to attempt to differentiate the problem variables essential in problem solving by the excellent student, the average student and the poor student.

The variables

All the mathematical problem variables defined in a previous paper (Lepik 1990) were analyzed separately for the three groups of students of different achievement levels mentioned above.

The definitions of the variables are as follows.

Linguistic variables

X_1- Space characters: The number of space characters in the problem;
X_2- Letters: The number of letters in the problem;
X_3- Words: The number of words in the problem;
X_4- Number of words, figures and units in the problem;
X_5- Mean word length: X_2/X_3;
X_6- Number of words of 6 or more letters;
X_7- Proportion of words of 6 or more letters: X_6/X_3;
X_8- Number of sentences in the problem;
X_9- Mean sentence length in space characters: X_1/X_8;
X_{10}- Mean sentence length in words: X_3/X_8.

Structural variables

The structural components of algebraic verbal problems are quantities and their mathematical relations. All quantities of the problem can be divided into three groups: known, calculable and auxiliary quantities. The mathematical operations between values required to solve the problem are determined by the formulation of the problem, and expressed by the formulas used and equations composed. To examine the logical structure of the problems the graphs representing all the values and their relations in the problem, were studied (Lepik 1990). The problem graphs can function as structural models of the problems. All values of the problem are denoted as nodes in the graph; all the formulas and equations are denoted as graph vertices. Graph nodes are connected with vertices by graph arcs. Each connection denotes the relation between a value and its formula or equation. All the structural variables are listed below:

X_{11}- The number of values given;
X_{12}- The number of calculable values;
X_{13}- The number of auxiliary values (values not given in the problem, but needed to solve it);
X_{14}- The number of values not given: $X_{12}+X_{13}$;
X_{15}- The whole number of values in the problem: $X_{11}+X_{14}$;
X_{16}- The number of formulas required;
X_{17}- The number of equations required;
X_{18}- The whole number of logical operations: $X_{16}+X_{17}$;
X_{19}- The whole number of structural components: $X_{15}+X_{18}$;
X_{20}- The number of relations between the values and operations;

X_{21}- The maximum number of relations per value;
X_{22}- The maximum number of relations per operation;
X_{23}- The average number of relations per value: X_{20}/X_{15};
X_{24}- The average number of relations per operation: X_{20}/X_{18};
X_{25}- The rank of the problem graph: X_{20}/X_{19};
X_{26}- The number of cycles (closed contours formed by arcs) in the graph.

In addition to the variables listed above, a variable integrating linguistic and structural features of the problem was used:

X_{27}- The average number of words in the problem per relation: X_3/X_{20}.

To determine the values of listed variables a standard solution algorithm was written for each problem tested and a problem graph was built. Most of the problems could be presented in a clear algorithm. Where two or more different algorithms could be constructed, the choice depended on how the problem was solved in the standard elementary-level textbook of mathematics.

Design

To start the study 100 verbal problems were pilot tested to determine their relative difficulty (in terms of the percentage of the correct answers). 35 problems representative for their level of difficulty were chosen. It is important to note that all the problems were similar to those presented in the standard textbooks and their solution had been drilled beforehand. The chosen problems were distributed among 6 tests of 5 or 6 problems. On the basis of the results of the pilot study the problems were arranged starting with the easiest ones in the order of growing difficulty. The typed-out tests were given to groups of students to be individually solved. The time for each task was not limited, but the whole testing session was invariably 45 minutes. All the test problems were solved by all the students in six sittings. The students were instructed to do all their work on the test sheets. To record the individual time spent on solving a problem an electrical digital timer was in good view in front of the classroom and the students were asked to write down the time when they started and finished each problem.

Subjects

Students aged 13-15 of five junior secondary schools were the subjects of this study. The 150 students were known to be of different levels in their mathematics progress.

To determine the achievement level of a student, the percentage of the correct solutions of the possible number was used. According to their scores all the students were grouped into "above average", "average" and "below average" groups. The "above average" groups included the students whose results were higher than P+0.5@; the "average" groups was made up of students who scored in the interval from P-0.5@ to P+0.5@; the "below average" group embraced the students whose results were lower than P-0.5@, where P is the mean percentage of the correct solutions and @ is its standard deviation.

There were 48 students in the "above average" group, 34 in the "average" group and 48 in the "below average" group.

Results

Performance variables

Two variables were used to measure performance:

(1) the average percentage of the correct answers;

(2) the average-time needed.

CORRECT ANWERS. A point was given for each logically correct presentation, even if the final answer was incorrect. As the test was designed to measure problem-solving skills and not computational ones, this approach seemed to be rational.

TIME. This variable measured the time spent in solving a problem. It did not define how much of the time was actually spent on task-solving and how much on other activities such as looking around or daydreaming.

Correlation Analysis

To find out the role of problem variables in guaranteeing a performance success, the correlation analysis was used. According to the aim of the paper the analysis was carried out separately on the results of the two contrasting groups of "above average" students and of "below average" students. The results of the analysis are summarized in Table 1.

As can be seen in Table 1 the linguistic variables were insignificant in predicting performance results in both groups. The only variable to be significant was X_{27}- the average number of words per each relation ("below averages": r = 0.38, "above averages": r = 0.49). It is the variable of information density in the wording of the problem. That means that a careful and more detailed verbal presentation of the problem will help to get a better insight into the mathematical connections and relations of the problem. However, most of the linguistic variables proved to be significantly correlated with the other performance variable - the time spent to solve the problem. Besides the correlation was regularly higher in the group of "above averages". The problem's textual readability being more significant for "above averages" seems to be an unexpected finding. Obviously, it does not measure reading difficulties, but indicates differences in the problem solving strategies in the two groups of students. "Above averages" seemed to be more skilled in and more attentive to the information in the wording of the mathematical problem. The students of above-average academic progress seem to be aware of the importance of the stage of information analysis and choice of working strategy in problem solving whereas the students of below-average progress underrate it. That is why the linguistic variables of the problem are better correlated with the time variables in the group of "above averages" and are not significantly correlated with the proportion of the correct answers. Maybe the point is that the "below averages" are more oriented to mechanical application of formulas they know and have used before than to the study of the underlying mathematical relations in the problem.

As can be seen in Table 1, the used structural variables appeared to be applicable in predicting both performance variables in both groups of students. X_{13} - the number of auxiliary quantities required in solving the problem correlated best with the proportion of correct answers (r = -0.51, r = -0.56). The best predictor of time variable was X_{12} - the number of computable quantities (r = 0.75, r = 0.55). The role of different logical operations (formulas, equations) manifests differences between the achievement groups. In the group of "below average" students X_{16} - the number of formulas needed in the solution is significantly correlated with the proportion of the correct answers (r = -0.33) but X_{17} - the

number of equations composed appeared insignificant (r = -0.12). In the group of "above average" students on the contrary, the number of applied formulas is not significant (r = -0.24) and the number of equations required is highly correlated with the proportion of the correct answers (r = -0.46). These results indicate that the "below average" group may solve the problem provided that only some known formulae are to be applied, but when it comes to making up equations then the problem proves too difficult for them. The students of below-average academic progress seem to be short of skills to synthesize and process information needed in the verbal presentation of the problem. Formula application is no problem in the advanced group and so the correlation between the number of formulas and the performance variables is low. The significant negative correlation between the number of equations in the solution and the proportion of the correct answers indicates that there are specific difficulties in information synthesizing in the "above average" group, too. But these difficulties (contrary to those in the "below average" group) can be overcome

Summary

Difficulties experienced by many students in solving routine problems seem to be related to their failure to identify the properties of the problem. To solve the problem one must identify the relationships between the data presented and choose a good strategy of processing and utilizing them. The understanding of the underlying mathematical relationships in a problem requires skills of identification of and operation with the different forms of expression of the relationships. This is difficult for low achievers. They usually have gaps in their knowledge of mathematics. Their operation skills are also unstable. So several pupils could choose and carry out the first operation but they were unable to proceed. The task of keeping the result of the first operation in mind and using it in the next step was beyond them.

The results of the study show that students of poor academic progress are often short of information processing and strategy-choosing skills. The skill of seeing isomorphism in problems of identical mathematical structure could be observed only in good problem solvers. On the contrary low achievers were unable to apply analogy in problem solving. They took each problem as a new one requiring original solution.

The study shows that low achievers need special instruction in problem solving, in analyzing the structural properties of problems and classifying them into types of problems.

Table 1
Correlation between the problem variables and the performance variables in student groups of different academic achievement.

Variables	Below average		Above average	
	Proportion of the correct answers	Time	Proportion of the correct answers	Time
X_1	-0.15	0.34*	0.07	0.45*
X_2	-0.14	0.33	0.05	0.43*
X_3	0.14	0.41*	0.09	0.51*
X_4	-0.13	0.40*	0.11	0.52*
X_5	-0.09	0.13	-0.17	-0.11
X_6	-0.22	0.13	-0.07	0.18
X_7	-0.20	0.19	-0.25	0.22
X_8	-0.21	0.36*	0.05	0.46*
X_9	0.02	0.16	0.11	0.24
X_{10}	0.05	0.28	0.15	0.23
X_{11}	0.03	0.31	0.15	0.38*
X_{12}	0.15	0.75*	0.08	0.55*
X_{13}	-0.51*	-0.25	-0.56*	-0.10
X_{14}	-0.45*	0.22	-0.55*	0.25
X_{15}	-0.31	0.38	-0.29	0.46*
X_{16}	-0.33*	-0.29	-0.12	-0.33*
X_{17}	-0.24	0.41*	-0.46*	0.46*
X_{18}	-0.47*	0.24	-0.56*	0.27
X_{19}	-0.41*	0.35*	-0.44*	0.41*
X_{20}	-0.40*	0.41*	-0.42*	0.44*
X_{21}	-0.34*	0.42*	-0.29	0.38*
X_{22}	-0.11	0.45*	-0.03	0.52*
X_{23}	-0.32	0.36*	0.17	0.36*
X_{24}	0.09	0.37*	0.17	0.45*
X_{25}	-0.31	0.42*	-0.29	0.43*
X_{26}	-0.30	0.47*	-0.30	0.44*
X_{27}	0.38*	-0.17	0.49*	-0.17

* $P < 0.05$

References

Austin J.D., Lee M.A.B. (1982). Readability and mathematics test item difficulty. *School Science and Mathematics 82, 284-290.*

Byers V., Erlwanger S. (1985). Memory in mathematical understanding. *Educational Studies in Mathematics 16, 259-282.*

Jerman M., Mirman S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics 5, 317-362.*

Krutetskii, V.A. (1976). *The psychology of mathematical abilities in schoolchildren.* Chicago, Univeristy of Chicago Press.

Lepik M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics 21, 83-90.*

Magne O. (1989). Mathematics learning of the handicapped student. *Zentralblatt für Didaktik der Mathematik, Januar.*

Moyer J.C., Moyer M.B., Soowder L. (1984). Story-problem formats: Verbal versus telegraphic. *Journal in Mathematic Education 15, 64-68.*

Nesher,P., Teubal E. (1975). Verbal cues as an interfering factor in verbal problem solving. *Educational Studies in Mathematics 6, 41-52.*

Pfeiffer K., Feinberg G., Gelberg G. (1987). Teaching productive problem solving attitudes. In: Berger, D.E., Pezdek, K., Banks, W.P. (eds.) *Applications of Cognitive Psychology: Problem Solving, Education and Computing.* London, Erlbaum: 99- 107.

Sochor A.M. (1974). *Logičeskaja struktura učebnogo materiala (*Logical structure of learning material). Moscow, Pedagogika.

Wright J.P., Wright C.D. (1986). Personalized verbal problems. *The Journal of Educational Research 79, 358-362.*

Glottometrika 14, 1993, 122-138

Comparison of Effectiveness of Various Basic Vocabularies

Haruko Sanada, Tokyo

0. Introduction

This paper is an attempt to analyze the effectiveness of basic vocabularies. "Effectiveness" can be defined as:

(a) an abstract meaning indicating how much semantic field a word covers;
(b) a concrete meaning indicating frequency of a word in a text.

In this paper (b) is measured.

Various lists of basic vocabularies have been published and this analysis is concerned with comparison of effectiveness of Japanese, English, French, and Spanish basic vocabularies in view of word coverage of a text.

As samples of basic vocabularies the following materials are used.

Japanese: 6,000 words ('Basic 6,000 words') including the 2,000 most basic words ('Basic 2,000 words') listed in *A Study of the Fundamental Vocabulary for Japanese Language Teaching* [7]. These words are selected from the point of view of foreign adults and include words of everyday life and abstract words and seem to reflect the vocabulary of modern Japanese society with very little bias.

'Basic 2,000 words' and 'Basic 6,000 words' have been obtained in the following way: word groups were first selected from *Word List by Semantic Principles* [1] by 22 specialists in Japanese language teaching, Japanese linguistics, and language education in order to "get a proper standard of general and basic Japanese vocabulary which foreign learners such as foreign students have to learn at the beginning as a basis for their further specialized field or occupational training" [7]. These word groups were modified in a second selection taking into account a bias in the first selection and defects of the data to obtain the final 'Basic 2,000 words' and 'Basic 6,000 words'.

French: about 5,000 words listed in *Dictionnaire du Vocabulaire essentiel* [2], which is used as data in *A contrastive study of the fundamental vocabulary of Japanese, German, French and Spanish* [8]. This dictionary was first published in French "[Notre dictionnaire], qui s'adresse surtout aux étrangers pourvus d'une certaine culture, mais dont les connaissances de français ne sont pas étendues" [10], and translated into Japanese "with a view to offer the first level and middle level learners a French-Japanese dictionary, which is useful and can cope with modern French sufficiently and essentially" [2]. This dictionary is used in this analysis since the original French version was compiled for foreign adult learners, and corresponds in its conception to 'Basic 2,000 words' and 'Basic 6,000 words'. This dictionary is also used as data in *A contrastive study of the fundamental vocabulary of Japanese, German, French and Spanish* [8], and will be useful material for my further study.

Spanish: about 5,000 words listed in *Diccionario de Vocabulario Fundamental del Español* [4], which is also used in *A contrastive study of the fundamental vocabulary of Japanese, German, French and Spanish* [8]. It is used in this study since it is issued in Japan "with the same reason and purpose" [4] as the above-mentioned *Dictionnaire du Vocabulaire essentiel* [2], and it has also about 5,000 words like the French one.

English: about 5,000 words listed in *The New Horizon Ladder Dictionary of the English Language* [5], which was translated into Japanese in the same series with *Dictionnaire du Vocabulaire essentiel* [2] and *Diccionario de Vocabulario Fundamental del Español* [4]. This dictionary was published in America "for readers of English as a second language" [9] as in the case of the French one, and translated into Japanese. 5,000 words of all the headwords in this dictionary are divided into five levels and each group of 1,000 words is marked with the frequency of their use from (1) to (5). In this study the 2,000 words of levels (1) and (2), and the 5,000 words of all levels are considered to correspond to each 'Basic 2,000 words' and 'Basic 6,000 words' of Japanese. However, in this dictionary there are included, under one headword, derivatives like noun, verb, adverb, and adjective, so the real number of listed words is estimated to be more than 5,000 words. In this study derivatives are regarded as the same words as the headwords marked with the frequency level.

As a sample text, *Declaration of the rights of the child* adopted by the United Nations in 1959 is used. The title, *Preamble*, from "Principle 1" to "Principle 10", and "Publicity to be given to the Declaration of the Rights of the Child" are analyzed. This document is selected as the sample text because various versions with the same contents are issued, and the texts in English, French,

and Spanish, which are official languages for the United Nations, are available in addition to the Japanese version.

In *A contrastive study of the fundamental vocabulary of Japanese, German, French and Spanish* [8], German is also compared, and *Deutscher Grundwortschatz* [3] (about 5,000 words) is published as a basic vocabulary in the same dictionary series, so an analysis of German was considered worth undertaking. However, the part of "Publicity to be given to the Declaration of the Rights of the Child" in the sample text has not been officially translated into German, and the analysis of the German version was abandoned.

1. Effectiveness of 'Basic 2,000 words' and 'Basic 6,000 words' in the Japanese version of "Declaration of the rights of the child"

Part of the Japanese version of "Declaration of the rights of the child" is shown in Example 1.

Example 1. Part of "Declaration of the rights of the child" (Japanese version)

第１条

児童は、この宣言に掲げるすべての権利を享有する。すべての児童は、いかなる例外もなく、自己又はその家族について、人種、皮膚の色、性、言語、宗教、政治上その他の意見、国民的もしくは社会的出身、財産、門地もしくは他の地位のため差別を受けることなく、平等に前記の権利を享有することができる。

One word in this study is counted in principle as one unit, but there are some exceptions:

- compound words in the text which are included in 'Basic 2,000 words' or 'Basic 6,000 words' are also counted as one unit. For example, the word "*sono hoka* (other)" which consists of "*sono*" and "*hoka*" is included in 'Basic 6,000 words', so it is counted as one word.

- for comparison of the Japanese version with the English, French, and Spanish ones, which will be discussed later, "Noun + *suru* (compound verb)" and "Noun + *da* or *na* (compound adjective)" are counted as one word, and the

part which is "Noun" in these words is compared with 'Basic 2,000 words' and 'Basic 6,000 words'.

For example, "*shuppan-suru* (publish)" and "*koufuku-da/-na* (happy)" are counted as one word since the words "*shuppan+suru*" and "*koufuku+da* or *na*" correspond to the following words: "publish (English)", "publier (French)", "publicar (Spanish)", and "happy (English)", "heureux (French)", "feliz (Spanish)". The words "*shuppan* (publication)" and "*koufuku* (happiness)" are confirmed to be included in 'Basic 2,000 words'.

- for comparison of the expressions in Japanese and European languages, phrases like "*oite* (at)", "*tsuite* (about)", "*toshite* (as)", "*atatte* (per)", and "*toiu* (called)" are counted as one word.

- prefixes and suffixes are included in 'Basic 2,000 words' and 'Basic 6,000 words', prefixes "*kaku-* (each)", "*dai-* (big)", and "*kou-* (wide)", suffixes "*-nen* (year)", "*-nado* (and so on)", "*-jou* (from the viewpoint of)", "*-kan* (feeling)", "*-teki* (adjective)", and "*-sei* (noun)" are counted as one word.

The numbers of different words and running words in the Japanese text are as follows:

	different words		running words	
- particles and auxiliary verbs	22 words	7.2%	402 words	34.0%
- up to 'Basic 2,000 words'	148 words	48.7%	849 words	71.9%
- up to 'Basic 6,000 words'	224 words	73.7%	1048 words	88.7%
- whole text	304 words	100.0%	1181 words	100.0%

This means that on the assumption of having knowledge of 'particles and auxiliary verbs', we can read 71.9% of this text with only 'Basic 2,000 words', and 88.7% with 'Basic 2,000 words' and 'Basic 6,000 words'. The number of particles and auxiliary verbs is only 22 (7.2%), but they cover 34.0%, that is a third of the running words.

A graph is drawn of the number of running words versus the number of different words (Figure 1) showing the cumulative frequency including particles and auxiliary verbs. The four arcs of the graph correspond to the word groups 'particles and auxiliary verbs', 'Basic 2,000 words', 'Basic 6,000 words', and 'others'. A regression curve of the form $Y = aX^b$ $(0 < b < 1)$ which crosses the points $(X = 0, Y = 0)$ and $(X = 100, Y = 100)$ is fitted to all data with the result

$a = 19.36$ and $b = 0.357$

using the method of least squares.

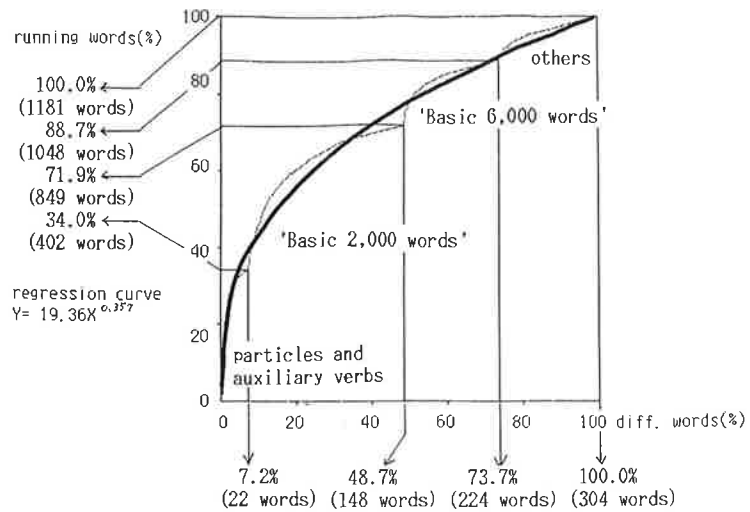


Figure 1. Effectiveness of Japanese basic vocabulary for "Declaration of the rights of the child"

The parameters a and b are obtained as follows (the same method is also applied to English, French, and Spanish versions which are analysed in later sections). From

(1) $Y = aX^b$,

we obtain by logarithmic transformation

(2) $\log Y = \log aX^b = \log a + b \log X$.

Let $y = \log Y$, $A = \log a$, and $x = \log X$, then (2) can be re-written as

(3) $y = A + bx$.

Evidently, (1) crosses the point ($X = 100$, $Y = 100$). If these values are substi-

tuted in (3), it will reduce to

(4) $\log 100 = A + b \log 100$.

Solving this for b yields

(5) $b = (\log 100 - A) / \log 100$.

Substituting (5) in (3) we obtain

$$(6)\quad y = A + \frac{(\log 100 - A)x}{\log 100} = A + \left(1 - \frac{A}{\log 100}\right)x.$$

The error between (6) and the observed values y_o is formulated as

$$(7)\quad e = y_o - y = y_o - A - \left(1 - \frac{A}{\log 100}\right)x.$$

To apply the method of least squares,

$$(8)\quad \langle e \rangle^2 = S = \sum_{i=1}^{n}\left[y_{oi} - A - \left(1 - \frac{A}{\log 100}\right)x_i\right]^2$$

should be minimized (n = number of data). Thus

$$(9)\quad \frac{dS}{dA} = 2\sum_{i=1}^{n}\left(-1 + \frac{x_i}{\log 100}\right)\left[y_{oi} - A - \left(1 - \frac{A}{\log 100}\right)x_i\right] = 0.$$

After performing the multiplications and summing in (9) we obtain

$$(10)\quad \sum_{i=1}^{n} y_{oi} - nA - \left(1 - \frac{A}{\log 100}\right)\sum_{i=1}^{n} x_i - \frac{1}{\log 100}\sum_{i=1}^{n} x_i y_{oi} + \frac{A}{\log 100}\sum_{i=1}^{n} x_i + \left(1 - \frac{A}{\log 100}\right)\frac{1}{\log 100}\sum_{i=1}^{n} x_i^2 = 0.$$

Multiplying (10) by $\log 100$ yields

$$\log 100 \sum_{i=1}^{n} y_{oi} - (\log 100)nA - (\log 100)\sum_{i=1}^{n} x_i + A\sum_{i=1}^{n} x_i$$

(11)

$$- \sum_{i=1}^{n} x_i y_{oi} + A\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2 - \frac{A}{\log 100}\sum_{i=1}^{n} x_i^2 = 0.$$

Solving (11) for A will yield

$$A = \frac{\log 100 \sum_{i=1}^{n} y_{oi} + \log 100 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_{oi} + \sum_{i=1}^{n} x_i^2}{\frac{1}{\log 100}\sum_{i=1}^{n} x_i^2 + n\log 100 - 2\sum_{i=1}^{n} x_i} \quad (12)$$

The parameter b is obtained from (5) and (12). The parameter a is obtained from (12) and $a = e^A$.

If the class of 'particles and auxiliary verbs' is excluded, the above table becomes:

	different words		running words	
- up to 'Basic 2,000 words'	126 words	44.7%	447 words	57.4%
- up to 'Basic 6,000 words'	202 words	71.6%	646 words	82.9%
- whole text	282 words	100.0%	779 words	100.0%

This means that even if 'particles and auxiliary verbs' are excluded, we can read 82.9% of this text with 202 words of 'Basic 2,000 words' and 'Basic 6,000 words'.

The point of covering 50% of the running words is at 35 different words (11.5%) with particles and auxiliary verbs, and it is at 72 different words (25.5%) without particles and auxiliary verbs. This difference seems to be due to the influence of particles and auxiliary verbs.

2. Effectiveness of basic vocabulary in the English version of "Declaration of the rights of the child"

Part of the English version of "Declaration of the rights of the child" is shown in Example 2.

Example 2. Part of "Declaration of the rights of the child" (English version)

PRINCIPLE 2

The child shall enjoy special protection, and shall be given opportunities and facilities, by law and by other means, to enable him to develop physically, mentally, morally, spiritually and socially in a healthy and normal manner and in conditions of freedom and dignity. In the enactment of laws for this purpose the best interests of the child shall be the paramount consideration.

For the analysis of the text, a space is used as word boundary. The article "an" is included in "a", and the conjugational forms and participles of a verb are included in its infinitive.

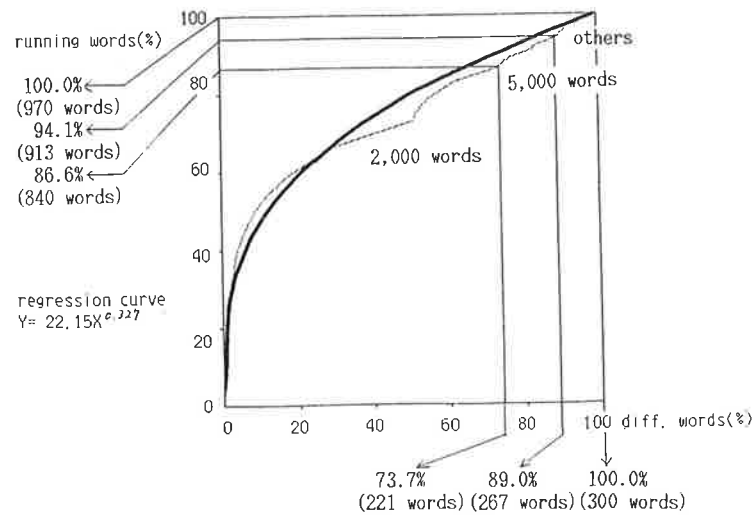


Figure 2. Effectiveness of the English basic vocabulary for "Declaration of the rights of the child"

The numbers of different words and running words in the English text are as follows:

	different words		running words	
- up to '2,000 words'	221 words	73.7%	840 words	86.6%
- up to '5,000 words'	267 words	89.0%	913 words	94.1%
- whole text	300 words	100.0%	970 words	100.0%

The number of running words is 970 which is 17.9% less than 1181 words of the Japanese version. In the analysis of the Japanese version, phrases like "*tsuite* (about)", "*atatte* (per)", and "*toiu* (called)", are counted as one word, so if they were divided into parts, the difference of the numbers of running words between Japanese and English would be much more. Though 'Basic 2,000 words' and 'particles and auxiliary verbs' cover only 71.9% of the running words, '2,000 words' covers 86.64% of the running words, and '5,000 words' covers 94.1% of the running words in the English version. However, it must be considered that '5,000 words' in the English basic vocabulary includes many derivatives.

A graph of the cumulative frequency of the number of running words versus the number of different words is drawn in Figure 2.

Comparing graphs between English (Figure 2) and Japanese (Figure 1), at the boundaries between '2,000 words', '5,000 words', and 'others', the line does not have pronounced kinks like the Japanese version, and it continues smoothly. Thus, part of 'Basic 6,000 words' and 'others' are also used frequently like 'Basic 2,000 words' in the Japanese version, but the range of '2,000 words', '5,000 words', and 'others' is in proportion to frequency. The expression for the regression curve of the graph for the English version is

$$Y = 22.15\, X^{0.327}\,.$$

Table 1
The regression curves for 4 languages ($1 \leq x \leq 10$)

x	$y = f(x)$ (regression curve)			
	Japanese	English	Spanish	French
1	19.36	22.15	25.43	28.62
2	24.79	27.79	31.25	34.55
3	28.64	31.73	35.25	38.58
4	31.74	34.87	38.40	41.71
5	34.36	37.51	41.03	44.32
6	36.67	39.82	43.32	46.57
7	38.74	41.88	45.35	48.56
8	40.63	43.75	47.19	50.35
9	42.38	45.47	48.87	51.99
10	44.00	47.06	50.43	53.50

Japanese: $y = 19.36x^{0.3565}$ English: $y = 22.15x^{0.3273}$
Spanish: $y = 25.43x^{0.2973}$ French: $y = 28.63x^{0.2716}$

The values of Y are always greater than the Japanese ones between $X = 1$ and $X = 10$ (see Table 1), and the differential values of this expression are also greater than those for Japanese until $X = 5$ at integral intervals (see Table 2). It can be seen from the table of differential values (Table 2) that the changing speed of values for the English version is faster than that for the Japanese one.

The point of covering 50% of the running words is at 27 different words (9.0%). It is less than 35 words (11.5%) in the Japanese version (including 'particles and auxiliary verbs'). The initial rise of the curve of accumulation of different words is steeper for English than for Japanese.

Table 2.
Differential values of regression curves for 4 languages ($1 \leq x \leq 10$)

x	dx/dy			
	Japanese	English	Spanish	French
1	6.90	7.25	7.56	7.78
2	4.42	4.55	4.65	4.69
3	3.04	3.46	3.49	3.49
4	2.83	2.85	2.85	2.83
5	2.45	2.46	2.44	2.41
6	2.18	2.17	2.15	2.11
7	1.97	1.96	1.93	1.88
8	1.81	1.79	1.75	1.71
9	1.68	1.65	1.61	1.57
10	1.57	1.54	1.50	1.45

Japanese: $dx/dy = 6.90x^{-0.6435}$ English: $dx/dy = 7.25x^{-0.6727}$
Spanish: $dx/dy = 7.56x^{-0.7027}$ French: $dx/dy = 7.78x^{-0.7284}$

3. Effectiveness of basic vocabulary in the French version of "Declaration of the rights of the child"

Part of the French version of "Declaration of the rights of the child" is shown in Example 3.

For the analysis of the text a space or an apostrophe are used as word boundary. The articles "une" and "la" are included in "un" and "le". The feminine form of an adjective is included in the masculine form, but singular and plural

forms are distinguished. The conjugational forms and participles of a verb are included in its infinitive. "au" and "du" are included in "à" and "de", "aux" and "des" are treated separately.

Example 3. Part of "Déclaration des droits de l'enfant" (French version)

PRINCIPE 3

L'enfant a droit, dès sa naissance, à un nom et à une nationalité.

The numbers of different words and running words in the French text are as follows:

	different words		running words	
- up to '5,000 words'	254 words	83.0%	1061 words	92.9%
- whole text	306 words	100.0%	1142 words	100.0%

The number of different words in the French version is 306, and the number of running words is 1142. Thus, the numbers obtained for the French text are close to those for the Japanese text. However, French '5,000 words' which contains fewer words than Japanese 'Basic 6,000 words' covers many more different words (83.0%) and running words (92.9%) than the Japanese one. This means that the Japanese different words depend on the 'others' vocabulary.

A graph of the cumulative frequency of the number of running words versus the number of different words is drawn in Figure 3. The line for French is similar to that for English, and it does not have a pronounced kink at the boundary between '5,000 words' and 'others', but continues more smoothly. The graph for the Japanese version on the other hand exhibits pronounced kinks. This means that low and high basicness are quite different with heavy weighting on the vocabulary of high basicness in the Japanese version. The graphs for the English and French versions are smoother because word frequency was the criterion for the selection of the respective basic vocabularies. The expression for the regression curve of the graph for the French version is

$$Y = 28.63\, X^{0.272}.$$

The values Y are always greater than the Japanese and English ones between $X = 1$ and $X = 10$ (see Table 1), and the differential values of this expression are also greater than those for Japanese and English until $X = 2$ at integral intervals (see Table 2). It can be seen from Table 2 that the changing speed of values for the French version is faster than that for the Japanese and English one.

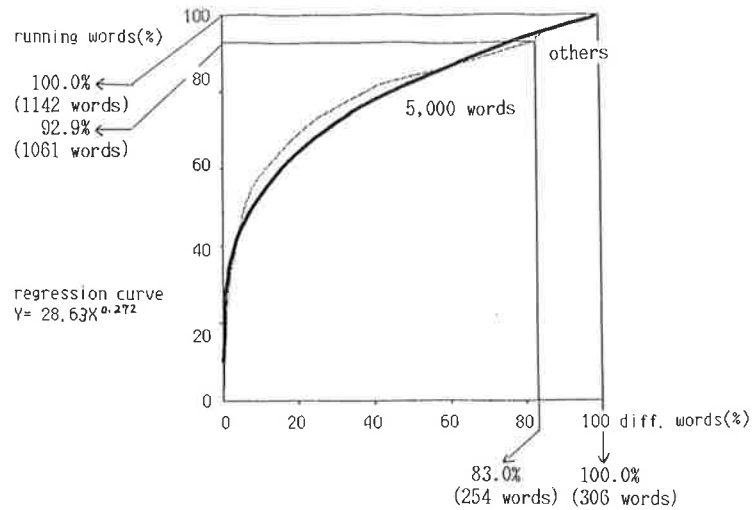


Figure 3. Effectiveness of the French basic vocabulary for "Déclaration des droits de l'enfant"

The point of covering 50% of the running words is at 19 different words (6.2%). This is much less than for the Japanese and English versions because the curve of accumulation of different words is much steeper for French than for Japanese and English.

4. Effectiveness of basic vocabulary in the Spanish version of "Declaration of the rights of the child"

Part of the Spanish version of "Declaration of the rights of the child" is shown in Example 4.

Example 4. Part of "Declaración de los derechos del nino" (Spanish version)

PRINCIPIO 4

El nino debe gozar de los beneficios de la seguridad social. Tendrá derecho a crecer y desarrollarse en buena salud; con este fin deberán proporcionarse, tanto a él como a su madre,

cuidados especiales, incluso atención prenatal y postnatal. El niño tendrá derecho a disfrutar de alimenación, vivienda, recreo y servicios médicos adecuados.

For the analysis of the text a space is used as the word boundary. The singular articles "una" and "la" are included in "un" and "el", the neuter article "lo" is also included in "el". The plural articles "unas" and "las" are included in "unos" and "los". The feminine form of an adjective is included in the masculine form, plural forms are also included in the singular. The forms like "algún" or "ningún" are included in the basic forms like "alguno" or "ninguno". The feminine, neuter, and plural forms of demonstrative pronouns or demonstrative adjectives are included in their masculine singular form. The conjugational forms and participles of a verb are included in its infinitive. Conjunctions "u" and "e", which are articulatorily modified, are included in "o" and "y". The infinitive or present participle with a personal pronoun like "separarse" is treated separately as "separar" and "se".

The numbers of different words and running words in the Spanish text are as follows:

	different words		running words	
- up to '5,000 words'	277 words	90.8%	968 words	95.2%
- whole text	305 words	100.0%	1017 words	100.0%

Characteristically the Spanish version with '5,000 words' covers such large fractions of different words (90.8%) and of running words (95.2%). This is clearly seen by comparing with the French version whose '5,000 words' covers 83.0% and 92.9%, respectively. Especially the high coverage of different words gives the impression that the whole text is easy to understand.

A graph of the cumulative frequency of the number of running words versus the number of different words is drawn in Figure 4.

The curve for Spanish is similar to English (Figure 2) and French (Figure 3). The graph does not have a pronounced kink as does the Japanese version (Figure 1). The expression for the regression curve of the graph for the Spanish version is

$$Y = 25.43\, X^{0.297}\,.$$

The values of Y are always greater than the English ones and smaller than the French ones between $X = 1$ and $X = 10$ (see Table 1), and the differential

values of this expression are located under those for French until $X = 2$ at integral intervals (see Table 2).

The point of covering 50% of the running words is at 22 different words (7.2%), thus also between English and French.

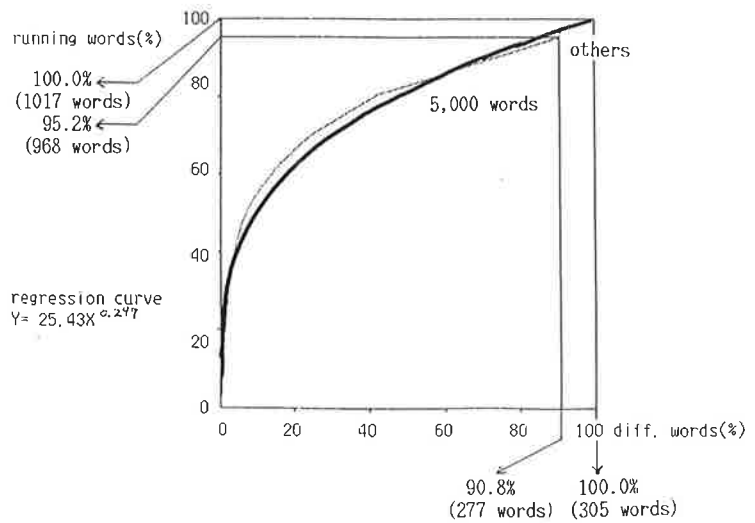


Figure 4. Effectiveness of the Spanish basic vocabulary for "Declaración de los derechos del nino"

5. Conclusions

It can be said that at least in the case of "Declaration of the rights of the child", English, French, and Spanish basic vocabularies with their '5,000 words' are more efficient than the Japanese ones with 'Basic 2,000 words' or 'Basic 6,000 words'.

The coverage of different words and running words of Japanese, English, French, and Spanish is shown in Table 3.

Graphs of regression curves for Japanese, English, French, and Spanish are shown in Figure 5. The values of the ordinates (the degree of cumulative

frequency) near the origin are highest for French, followed by Spanish, English, and Japanese. The curves for French and Spanish lie above those for English and Japanese.

Table 3
Coverage of four languages for "Declaration of the rights of the child"
(% different words / % running words)

	Japanese	English	French	Spanish
particles and auxiliary verbs	7.2 / 34.0	- / -	- / -	- / -
up to 2,000 words	48.7 / 71.9	73.7 / 86.6	- / -	- / -
up to 5,000 words	- / -	89.0 / 94.1	83.0 / 92.9	90.8 / 95.2
up to 6,000 words	73.7 / 88.7	- / -	- / -	- / -

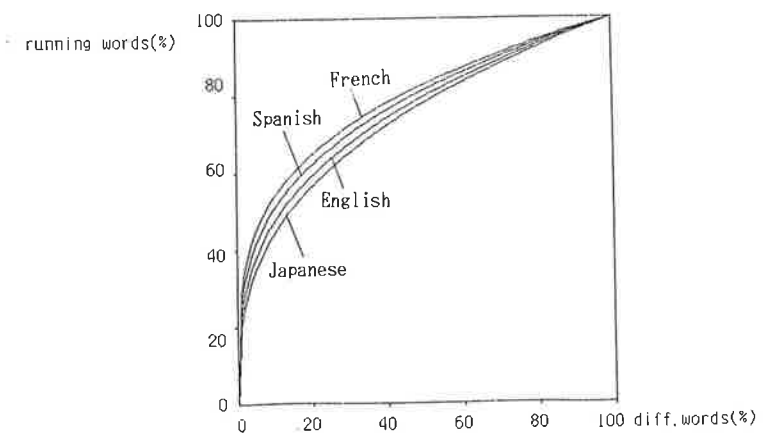


Figure 5. Regression curves for "Declaration of the rights of the child"

Hiroshi Nakano analysed the frequency of six languages [6] using 'The Little Prince', and he shows that the decreasing order is Japanese (including particles and auxiliary verbs), French, English, German, and Japanese (excluding particles and auxiliary verbs). The result is explained as follows:

". . . the French is on the top . . . up to number 10 in the frequency order. Thereafter the Japanese with auxiliary verbs and postpositional words moves to the top. Though the rank variation for other languages can not be expressed as a number, the corresponding postpositional words are counted in Japanese. The curve for Japanese will fall to the lowest if auxiliary verbs and postpositional words are removed from it. This is due to the fact that the preposition and auxiliary verbs for all other languages have been counted, whereas the corresponding postpositional words and auxiliary verbs are excluded from the Japanese" [6].

However, in this analysis of the text "Declaration of the rights of the child", the result is that the French basic vocabulary '5,000 words' covers many more different words and running words than the Japanese basic vocabulary 'Basic 6,000 words' including particles and auxiliary verbs. One of the reasons for this difference between these two studies is due to counting "Noun + suru (compound verb)" as one word in the Japanese version.

In this study it is observed that for an understanding of the sample text the Japanese basic vocabulary is not sufficient compared with foreign ones. This is probably due to the fact that Japanese has a unique structure of the vocabulary in which original Japanese words, loan words from China, and loan words from Europe constitute the same semantic field.

References

1. **The National Language Research Institute** (1964). *Bunrui goi hyo* (Word list by semantic principles). Tokyo, Shueishuppan.
2. **Matoré, G.** (trans. and ed. Nomura, J., Namekawa, A.) (1967). *Furansu kihongo 5000 jiten.* (Dictionnaire du vocabulaire essentiel). Tokyo, Hakusuisha.
3. **Iwasaki, E., Hayakawa, T., Koyasu, M., Hirao, K., Tetsuno, Y.** (1971). *Doitsu kihongo 5000 jiten* (Deutscher Grundwortschatz). Tokyo, Hakusuisha.
4. **Takahashi, M., Uritani, R., Miyagi, N., Contreras, E.** (1972). *Supein kihongo 5000 jiten.* (Diccionario de Vocabulario Fundamental del Español). Tokyo, Hakusuisha.
5. **Shaw, J.R.** (trans. and ed. Fukuda, R.) (1972). *Radaa eiwa kihongo 5000 jiten.* (The New Horizon Ladder Dictionary). Tokyo, Hakusuisha.
6. **Nakano, H.** (1976). "Hoshi no oujisama" 6-kakokugoban no goironteki kenkyu. *Keiryokokugogaku 79, 18-31* (Mathematical Linguistics No.79),

(translated version: A lexical survey of six versions of "The Little Prince" in various languages. In: Mizutani, Sh. (ed.), *Japanese Quantitative Linguistics*. Bochum, Brockmeyer 1989: 90-106).

7. **The National Language Research Institute** (1984). *Nihongo Kyoiku no tame no kihon goi chousa.* (A study of the fundamental vocabulary for Japanese language teaching). Research Report 78, Tokyo, Shueishuppan.
8. **The National Language Research Institute** (1986). *Nichi doku futsu sai kihon goi taishouhyo.* (A contrastive study of the fundamental vocabulary of Japanese, German, French and Spanish). Research Report 88, Tokyo, Shueishuppan.
9. **Shaw, J.R., Shaw, S. J.** (1990). *The New Horizon Ladder Dictionary of the English Language*. Revised and enlarged edition. New York, Penguin Books.
10. **Matoré, G.** (1963). *Dictionnaire du vocabulaire essentiel (les 5000 mots fondamentaux)*. Paris, Larousse.

Glottometrika 14, 1993, 139-160

Typological Indices and Language Classes:
A Quantitative Study

George Silnitsky, Smolensk

This paper, which is a further elaboration of Silnitsky & Jachontov (1986), is dedicated to a comparative study in quantitative terms of the following 31 languages:

ARAB - Arabic	INDO - Indonesian	O.CH - Old Chinese	TEL - Telugu
BURM - Burmese	JAP - Japanese	PERS - Persian	THAI - Thai
CHIN - Chinese	KHM - Khmer	RUS - Russian	TIB - Tibetan
CHUK - Chukchee	KOR - Korean	SAN - Sanskrit	TUR - Turkish
ENG - English	MANC - Manchurian	SWA - Swahili	UR - Urdu
FRE - French	MANI - Maninka	TAD - Tadzhik	VIET - Vietnamese
GER - German	MARI - Mari	TAG - Tagalog	YID - Yiddish
HIN - Hindi	MON - Mongolian	TAN - Tangut	

The properties of these languages are evaluated in terms of a set of indices proposed by J.H. Greenberg (1960) with certain modifications (cf. Krupa 1965):

(1) *SYN - Synthetical index:* relation of the number of morphs to the number of words (M/W).

(2) *AGGL - Agglutinational index:* relation of the number of agglutinative constructions to the number of morph sutures (A/Su).

(3) *COMP - Compositional index:* relation of the number of roots to the number of words (R/W).

(4) *DER - Derivational index:* relation of the number of derivational morphemes to the number of words (D/W).

(5) *INF - Inflectional index:* relation of the number of inflectional morphemes to the number of words (I/W).

(6) *PREF - Prefixal index:* relation of the number of prefixes to the number of words (P/W).

(7) *SUF - Suffixal index:* relation of the number of suffixes to the number of words (S/W).

(8) *ISOL - Isolational index:* relation of the number of non-inflected words, whose syntactic function is determined solely through word-order, to the number of words (O/W).

(9) *PI - Index of pure inflection:* relation of the number of inflected noncon-cordant words to the number of words (Pi/W).

(10) *CONC - Index of concordance:* relation of the number of concordant words to the number of words (Co/W).

The last three indices have in the position of denominator not the number of nexus connections, as in Greenberg, but the number of words, which gives us a common basis for a more systematic comparison of the data pertaining to the majority of the indices under consideration.

The eleventh "analytical" index (ANAL) is introduced defined as the relation of the number of auxiliary (delexicalized) words to the total number of words (Aux/W). The values of the indices for every language have been determined on texts of approximately 100 words. The indices for the first 26 languages of our list were borrowed from Kasevič & Jachontov (1982); the indices for the last 5 languages were calculated by S. Je. Jachontov. "It has been considered possible to dispense with a mathematical substantiation of the sample size, a calculation of the significance level, etc. As may be seen from Greenberg 1960, Cowgill 1963, Krupa 1965, Krupa & Altmann 1966, Pierce 1966, and others, the values of the indices are sufficiently stable and do not vary significantly with text samples of different magnitudes" (Kasevič & Jachontov 1982:6).

The samples were taken from various narrative genres: modern fiction, fairy tales, memoirs, newspaper articles, etc.

The values of the 11 typological indices for the 31 languages considered in this paper are given in Table 1.

These initial data are processed by means of various statistical procedures (correlational analysis, factor analysis and others) with the aim of establishing

(1) a classification of typological indices,
(2) a classification of languages,
(3) a selection of the most diagnostically relevant indices.

Table 1
Typological indices for 31 languages

	1	2	3	4	5	6	7	8	9	10	11
VIET	1.46	1.00	1.34	.02	.08	.02	1.00	1.00	.00	.00	.14
KHM	1.50	.98	1.34	.04	.14	.12	.04	.90	.10	.00	.17
THAI	1.46	1.00	1.29	.04	.08	.06	.06	.95	.05	.00	.24
O.CH	1.26	1.00	1.19	.02	.06	.04	.04	.99	.01	.00	.12
CHIN	1.56	1.00	1.22	.04	.28	.00	.31	.99	.01	.00	.14
BURM	1.73	.93	1.21	.12	.39	.08	.41	.74	.27	.00	.26
TAN	1.32	.72	1.13	.01	.17	.09	.07	.93	.06	.01	.12
TIB	1.75	.98	1.30	.29	.16	.05	.39	1.00	.00	.00	.26
MANI	1.16	1.00	1.03	.06	.07	.08	.05	1.00	0.00	.00	.33
INDO	1.50	.99	1.18	.17	.08	.17	.06	.98	.00	.02	.26
TAG	1.42	1.00	1.02	.12	.22	.22	.08	1.00	.00	.00	.51
PERS	1.67	.83	1.10	.04	.53	.07	.50	.66	.03	.09	.30
TAD	1.67	.84	1.09	.06	.52	.06	.53	.75	.03	.12	.38
ENG	1.30	.57	1.03	.09	.18	.00	.27	.77	.19	.04	.37
FRE	1.54	.39	1.01	.12	.43	.19	.35	.59	.04	.37	.46
YID	2.00	.18	1.05	.17	.78	.16	.52	.79	.00	.15	.26
GER	1.57	.17	1.08	.07	.42	.03	.44	.40	.32	.28	.32
HIN	1.70	.26	1.07	.10	.52	.10	.60	.30	.40	.30	.40
UR	1.68	.32	1.04	.14	.50	.02	.62	.40	.34	.26	.38
ARAB	3.14	.50	1.00	.20	1.94	.34	1.13	.33	.40	.27	.26
SAN	2.60	.22	1.12	.53	.94	.22	1.22	.24	.46	.30	.21
RUS	2.11	.13	1.00	.38	.74	.17	.93	.31	.40	.29	.24
MANC	1.85	.96	1.04	.13	.68	.00	.81	.48	.52	.00	.02
MON	2.10	.97	1.10	.23	.77	.00	1.01	.44	.56	.00	.03
TUR	2.15	.93	1.01	.13	1.01	.00	1.14	.14	.60	.26	.01
MARI	2.02	.69	1.01	.18	.83	.00	1.01	.27	.47	.22	.11
KOR	2.31	.93	1.26	.04	1.01	.00	1.05	.26	.74	.00	.09
JAP	2.71	.86	1.29	.15	1.23	.00	1.38	.28	.72	.00	.07
TEL	2.61	.83	1.14	.48	.98	.04	1.42	.20	.64	.16	.18
CHUK	2.33	.81	1.06	.11	1.17	.21	1.02	.36	.29	.35	.07
SWA	2.51	1.00	1.03	.28	1.24	.94	.56	.62	.00	.38	.21

1. Classification of typological indices

The correlations between the 11 typological indices are represented in Table 2. The critical value of r with 29 degrees of freedom and $\alpha = 0.01$ is $|0.45|$. On this criterion our indices may be grouped into two "polar" clusters.

Cluster A encompasses the following 6 indices: SUF, SYN, INF, PI, DER, CONC. The first four indices are positively interconnected with one another and thus constitute the "nucleus" of the cluster. The remaining two indices, DER, CONC, are positively correlated with some, but not with all the nuclear indices and occupy a "peripheral" position in the cluster.

Cluster B includes 3 indices: AGGL, ISOL, COMP. AGGL, the nuclear element of the cluster, is positively correlated with the two peripheral indices, ISOL and COMP.

Let the **medial intraclusteral coefficient** (MIC) of an element represent the arithmetic mean of its correlations with all the other elements of the same cluster; the MIC of an index thus shows the degree of its "centrality" in the cluster.

Cluster A		Cluster B	
SUF:	.72	AGGL:	.45
SYN:	.70	ISOL:	.40
INF:	.67	COMP:	.40
PI:	.54		
DER:	.49		
CONC:	.41		

The **average intraclusteral coefficient** (AIC), defined as the arithmetic mean of the MIC's of all its elements, manifests the "degree of internal cohesion" of the cluster. The value of AIC is .59 for cluster A and .42 for cluster B.

The **medial extraclusteral coefficient** (MEC) of an element as regards some other cluster is the arithmetic mean of the correlations of this element with all the elements of the alien cluster. The MEC's of the indices of clusters A and B in relation to one another are given below.

Table 2
Correlations between the 11 typological indices

	AGGL	COMP	DER	INF	PREF	SUF	ISOL	PI	CONC	ANAL
SYN	-.22	-.11	.61	.88	.27	.86	-.74	.63	.49	-.35
AGGL	x	.45	-.32	-.23	.00	-.26	.45	-.18	-.65	-.38
COMP		x	-.21	-.32	-.23	-.22	.34	-.11	-.58	-.32
DER			x	.44	.30	.60	-.47	.37	.41	-.04
INF				x	.37	.85	-.77	.63	.54	-.32
PREF					x	-.02	.04	-.29	.46	.14
SUF						x	-.91	.88	.43	-.42
ISOL							x	-.88	-.62	.27
PI								x	.19	-.47
CONC									x	.20

Cluster A		Cluster B	
SUF:	-.46	AGGL:	-.31
SYN:	-.36	ISOL:	-.73
INF:	-.44	COMP:	-.26
PI:	-.37		
DER:	-.32		
CONC:	-.62		

The main opposition is between CONC, SUF and INF in the first cluster and ISOL in the second.

The arithmetic mean of all the correlations of the elements of one cluster with the elements of another constitutes the **average extraclusteral coefficient** (AEC) for these two clusters. The AEC for cluster A and B is -.42.

The MEC's of the remaining isolated indices PREF and ANAL as regards clusters A and B are as follows:

Cluster A		Cluster B
PRE:	.18	-.07
ANAL:	-.23	-.14

A more differentiated grouping of indices is effected by means of factor analysis as presented in Table 3.

The first of the five factors represented above serves to delimit cluster A and to differentiate it into "nuclear" and "peripheral" indices, exerting a stronger (negative) influence on the former (SUF, SYN, INF, PI) than on the latter (DER, CONC). This stratification corresponds closely enough to the ranking effected above on the MIC-principle; on both criteria SUF figures as the dominant, most representative element of the cluster.

The second factor singles out cluster B (AGGL, ISOL, COMP) by its positive influence and AGGL as the dominant element of the cluster.

Table 3
Factor analysis of typological indices

	SUF	SYN	INF	PI	DER	CONC	AGGL	ISOL	COMP	PREF	ANAL
F1	-.93	-.85	-.85	-.89	-.36	-.35	.13	.89	.12	.04	.62
F2	-.10	-.10	-.11	-.03	-.17	-.67	.93	.34	.30	-.02	-.50
F3	.02	-.33	-.40	.34	-.18	-.46	-.07	-.07	.16	-.96	-.04
F4	-.29	-.30	-.08	-.13	-.88	-.04	.13	.11	.02	-.15	-.17
F5	.10	.08	.15	.05	.04	.34	-.17	-.21	-.92	.13	.33

F3 deals in like fashion with PREF.

The remaining index, ANAL, is not marked out uniquely on any factor level: its isolated, unrepresentative status in the system thus finds additional corroboration.

The conclusion may therefore be drawn that the proposed classification of typological indices into two clusters with two isolated indices (PREF, ANAL) is sufficiently well-grounded.

This classification may be compared with that in Altmann & Lehfeldt (1973). The majority of the indices here have as their denominator not the number of words, as in our case, but the number of morphemes or syntactic (nexus) connections. Only six languages (Sanskrit, Persian, English, Swahili, Vietnamese, Turkish) are common to both lists.

All the more representative, in spite of these dissimilarities in initial criteria and linguistic material, is the relatively high degree of congruence between the results of the two investigations. Thus, the indices SUF, INF, PI, DER (and, to

a lesser degree, CONC and PREF) in Altmann & Lehfeldt are positively interconnected with one another. SYN is negatively correlated with SUF, DER, CONC, and others; but seeing that the relation of the number of words and morphemes is given here in inverse proportion as compared with our treatment and that of Greenberg (W/M instead of M/W) there is no actual divergence between the two classifications. Our cluster A (with the inclusion of the more loosely connected index PREF) finds its correlate in Altmann & Lehfeldt. These indices are negatively opposed to ISOL, which likewise agrees with our scheme.

On the other hand, the indices AGGL and COMP in Altmann & Lehfeldt are isolated from the others whereas in our classification they are positively connected with ISOL, constituting cluster B.

It therefore seems probable on the whole that the crucial correlations between the main typological indices are more or less constant on a wide variety of languages and may thus be considered as candidates for the status of relational universals.

2. Classification of languages

The correlational coefficients between the 31 languages listed above, calculated on the basis of the 11 typological indices, are presented in Table 4. The critical value of r is $|.78|$.

We shall provisionally regard language classes as hierarchically stratified into three levels:

(a) The "nuclear" level of a class encompasses languages whose mutual correlations are not lower than .93.

(b) The "peripheral" level of a class includes languages whose mutual correlations with one another and with the nuclear languages are not lower than .89.

(c) Languages whose mutual correlations with one another and with the languages of the preceding two levels are not lower than .78 constitute the "marginal" level of the corresponding class.

The nuclear level is obligatory for any language class; the other two levels are optional. Classes are "stratified" or "nonstratified" depending on whether they contain languages pertaining to the peripheral and/or marginal levels. The mini-

Table 4
Correlations between 31 languages

	VIET	KHM	THAI	0.CH	INDO	MANI	TAN	TAG
KHM	.99							
THAI	.99	.99						
O.CH	.99	.99	.99					
INDO	.99	.99	.99	.99				
MANI	.98	.98	.99	.99	.99			
TAN	.97	.96	.96	.98	.96	.98		
TAG	.96	.96	.97	.95	.98	.98	.94	
TIB	.94	.93	.95	.94	.95	.93	.90	.92
CHIN	.97	.96	.97	.97	.95	.95	.95	.93
BURM	.90	.92	.91	.91	.89	.90	.90	.89
PERS	.82	.83	.83	.83	.81	.84	.85	.84
TAD	.83	.83	.84	.84	.82	.85	.85	.86
ENG	.86	.87	.89	.90	.87	.93	.94	.90
FRE	.68	.69	.71	.72	.72	.78	.81	.80
YID	.46	.45	.45	.48	.46	.50	.61	.53
GER	.39	.40	.41	.45	.38	.50	.59	.45
HIND	.24	.27	.28	.31	.24	.35	.42	.32
UR	.37	.39	.41	.44	.37	.47	.54	.44
RUS	-.07	-.07	-.06	-.02	-.07	.00	.12	-.02
SAN	-.25	-.24	-.25	-.21	-.26	-.22	-.12	-.23
ARAB	-.10	-.09	-.12	-.10	-.15	-.12	-.01	-.07
CHUK	.21	.23	.20	.22	.16	.18	.26	.19
MANC	.55	.58	.56	.57	.51	.53	.58	.49
MON	.41	.44	.42	.43	.36	.38	.43	.35
TUR	.16	.20	.17	.19	.11	.15	.21	.12
MARI	.30	.30	.33	.33	.25	.32	.39	.33
KOR	.21	.25	.22	.22	.14	.17	.23	.16
JAP	.06	.09	.07	.07	.00	.02	.09	.02
TEL	-.02	.01	.00	.01	-.04	-.03	.02	-.04
SWA	.36	.36	.32	.33	.35	.30	.35	.38

Table 4. Continuation

	TIB	CHIN	BURM	PERS	TAD	ENG	FRE	YID
CHIN	.97							
BURM	.92	.96						
PERS	.85	.92	.94					
TAD	.88	.93	.94	.99				
ENG	.85	.88	.88	.86	.87			
FRE	.69	.74	.77	.86	.86	.90		
YID	.52	.60	.48	.74	.74	.66	.84	
GERM	.38	.48	.56	.66	.65	.76	.82	.76
HIN	.28	.39	.52	.63	.62	.64	.73	.69
UR	.43	.52	.63	.74	.72	.74	.81	.76
RUS	.03	.13	.26	.41	.38	.30	.50	.69
SAN	-.06	-.02	.12	.27	.23	.04	.25	.52
ARAB	-.03	.10	.22	.40	.36	.03	.30	.61
CHUK	.29	.42	.52	.67	.62	.28	.43	.62
MANC	.62	.70	.83	.81	.77	.61	.57	.54
MON	.52	.59	.73	.73	.69	.48	.46	.49
TUR	.26	.37	.53	.40	.54	.28	.34	.43
MARI	.49	.58	.68	.81	.82	.55	.63	.76
KOR	.29	.40	.58	.59	.54	.30	.32	.39
JAP	.19	.28	.45	.50	.46	.17	.24	.41
TEL	.18	.22	.39	.44	.41	.13	.21	.35
SWA	.34	.44	.44	.58	.54	.20	.43	.55

Table 4. Continuation

	GERM	HIN	UR	RUS	SAN	ARAB	CHUK	MANC
HIN	.96							
UR	.96	.98						
RUS	.74	.83	.82					
SAN	.48	.63	.62	.94				
ARAB	.39	.52	.50	.72	.79			
CHUK	.50	.62	.64	.72	.75	.88		
MANC	.58	.66	.71	.59	.55	.57	.82	
MON	.50	.62	.67	.64	.66	.63	.86	.98
TUR	.50	.67	.67	.72	.76	.76	.93	.90
MARI	.72	.85	.86	.79	.76	.71	.92	.92
KOR	.44	.60	.60	.62	.68	.74	.88	.91
JAP	.37	.56	.56	.69	.80	.81	.89	.84
TEL	.36	.56	.57	.75	.86	.73	.84	.80
SWA	.15	.16	.22	.27	.28	.65	.70	.47

Table 4. Continuation

	MON	TUR	MARI	KOR	JAP	TEL
TUR	.94					
MARI	.95	.98				
KOR	.95	.96	.94			
JAP	.91	.94	.90	.97		
TEL	.90	.93	.90	.91	.96	
SWA	.44	.46	.32	.42	.41	.31

mal value .78 is thus a necessary, but not a sufficient, condition for a correlation to fulfil a class-forming function. Correlations not lower than .78 will be further termed "critical".

Languages answering the above-mentioned class-forming conditions (i.e. figuring as elements of a certain class) will be regarded as **class-bound**. Four language classes may be discriminated in our material on these criteria.

Two of these classes are stratified:

(1) **Class I** (based mainly on classical isolating languages):

(a) *Nuclear level:* Vietnamese, Khmer, Thai, Old Chinese, Chinese, Indonesian, Maninka, Tangut, Tagalog;
(b) *Peripheral level:* Tibetan, Burmese;
(c) *Marginal level:* Persian, Tadzhik, English.

(2) **Class A** (based upon agglutinative languages):

(a) *Nuclear level:* Mongolian, Turkish, Mari, Korean;
(b) *Peripheral level:* Japanese, Telugu;
(c) *Marginal level:* Manchurian, Chukchee.

The remaining two classes are nonstratified (within the set of languages considered):

(3) **Class F** (markedly inflective languages): Sanskrit, Russian.

(4) **Class GH** (German, Hindi, Urdu) is the most heterogeneous (from the traditional typological point of view); we find it difficult at the present stage of our investigation to bring it down to a common typological denominator.

Two class-bound languages, belonging to different classes and positively correlated with one another, will be said to perform a "linking" function as regards these two classes. A full list of pairs of such "linking" languages is given below:

Burmese (I) - Manchurian (A)
Persian (I) - Manchurian (A)
Persian (I) - Mari (A)
Tadzhik (I) - Mari (A)
Russian (F) - Mari (A)
Mari (A) - Hindi (GH)
Mari (A) - Urdu (GH)
Hindi (GH) - Russian (F)
Urdu (GH) - Russian (F)
Telugu (A) - Sanskrit (F).

Languages not included into any class on the above-given criteria will be considered **class-free.**

Class-free languages are characterized by various degrees of cohesion with other languages:

(1) A "satellite" class-free language is critically correlated with *all* the languages of one of the three hierarchical levels of a certain language class (usually with all the marginal languages). Thus, French is significantly correlated with all three I-marginal languages (English, Persian, Tadzhik) and is therefore an I-satellite language.

(2) An "appended" language is critically correlated with some (at least one) but not all the languages of a hierarchical level (or levels) of a certain language class. Arabic is critically correlated with Japanese and Chukchee, on the one hand, and with Sanskrit, on the other, and is thus an A-appended and a F-appended language.

Satellite and appended languages, critically correlated with representatives of different language classes, perform a "connecting" function as regards these languages. Thus, French interconnects classes I (Tangut, Persian, Tadzhik, Tagalog, Maninka, English) and GH (German, Urdu).

(3) A "contingent" language is critically correlated with one or more class-free, but with no class-bound languages. Yiddish answers this condition, being critically connected with French. (It should be noted that Yiddish and French do not constitute a separate class on their own account since they lack a common nucleus.)

As in the case of typological indices considered above, individual languages are characterized by a set of generalized correlation coefficients:

(a) The medial intraclusteral coefficient (MIC) of a class-bound language, defined as the mean of its correlations with all the other languages of the same class, reflects its "degree of centrality" in the corresponding class.

(b) The medial extraclusteral coefficient (MEC) of any language (either class-bound or class-free) as regards an alien class is calculated as the mean of the correlations of this language with all the languages of the corresponding class thus representing the typological "distance" ("degree of proximity") of the former in respect of the latter.

Two new correlational criteria are to be introduced at this stage:

(c) The **general medial extraclusteral coefficient (GMEC)** of a language is the mean of all its MEC's, thus showing the degree of "centrality" of this language as regards the system of language classes external to it.

(d) The **general coefficient (GC)** of a language is the mean of its coefficients with all the languages of the set considered, thus illustrating the degree of its "centrality" in the whole system.

These correlational characteristics of our 31 languages are given in Table 5. We see that of the class-free languages French is mainly "oriented" towards classes I and GH, Arabic towards classes F and A, Yiddish towards class GH; Swahili occupies an approximately equally remote position as regards all four classes.

The criteria of class-oriented (GMEC) and general (GC) "centrality" do not give identical results. On the first of these the most central languages are Mari, Urdu, Manchurian and Yiddish, the most marginal Indonesian and Vietnamese, while according to the second the corresponding representatives are Persian, Tadzhik and Burmese, on the one hand, and Sanskrit, Arabic and Swahili, on the other.

Whole language classes are characterized by the following correlational criteria:

(a) The average intraclusteral coefficient (AIC) of a class is the mean of the correlations between all the languages of this class and shows its **degree of internal cohesion**.

(b) The average extraclusteral coefficient (AEC) of a class is the mean of all the mutual MEC's between these two classes and represents the typological **distance** separating them.

(c) The general average extraclusteral coefficient (GAEC) of a class, calculated as the mean of all its AEC's with the other classes of the system, reflects its generalized degree of **centrality** in the system.

Table 6 represents these generalized correlational interrelations between our four language classes.

Table 5
Correlational characteristics of 31 languages

	MIC	MEC				GMEC	GC
CLASS I		I	A	F	GH		
Vietnamese	.94		.24	-.16	.33	.14	.64
Khmer	.94		.26	-.16	.37	.16	.55
Thai	.94		.25	-.16	.38	.16	.54
O.Chinese	.94		.26	-.12	.40	.18	.56
Indonesian	.94		.19	-.17	.33	.12	.52
Maninka	.95		.22	-.11	.44	.18	.55
Tangut	.93		.28	.00	.52	.27	.59
Tagalog	.93		.20	-.13	.40	.16	.54
Tibetan	.92		.36	.00	.36	.24	.58
Chinese	.95		.45	.06	.46	.32	.64
Burmese	.91		.59	.19	.57	.45	.69
Persian	.86		.62	.34	.68	.55	.72
Tadzhik	.87		.61	.31	.66	.53	.71
English	.88		.35	.17	.71	.41	.62
CLASS A							
Mongolian	.93	.50		.65	.60	.58	.62
Turkish	.94	.27		.74	.61	.54	.52
Mari	.93	.46		.78	.81	.68	.65
Korean	.93	.31		.65	.55	.50	.52
Japanese	.92	.18		.75	.50	.48	.46
Telugu	.89	.12		.81	.50	.48	.42
Manchurian	.88	.62		.57	.65	.61	.65
Chukchee	.88	.32		.74	.59	.55	.55
CLASS F							
Sanskrit	.94	-.09	.73		.58	.41	.30
Russian	.94	.08	.69		.80	.52	.41
CLASS GH							
German	.96	.51	.50	.61		.54	.54
Hindi	.97	.40	.64	.73		.59	.54
Urdu	.97	.51	.66	.72		.63	.61
CLASS-FREE							
French		.77	.40	.38	.79	.58	.62
Arabic		.00	.73	.76	.47	.49	.35
Yiddish		.55	.50	.61	.74	.60	.57
Swahili		.38	.44	.28	.18	.32	.38

Table 6
Generalized correlational interrelations

	AIC	AEC				GAEC
		I	A	F	GH	
Class I	.92		.35	.00	.47	.27
Class A	.81	.35		.71	.60	.55
Class F	.94	.00	.71		.69	.47
Class GH	.97	.47	.60	.69		.58

The closest connections are between classes A - F and F - GH; the greatest typological distance separates classes I and F.

Class GH, characterized by the most "amorphous" typological structure, is at the same time the most "central" in the system. Class I, on the other hand, occupies the most marginal position.

Several variants of factor analysis (from 2 to 6 factors) were applied to our list of languages. Two of these, FA 4 (4 factors) and FA 6 (6 factors) are presented in Table 7.

These data give the following factorial corroboration to our four language classes:

CLASS I

FA 4: (1) F2: Diapason (-.63, .00) (includes French).
(2) F3: Diapason (-.19, .00). Exception: Tangut has value .19, which brings it close to Class F.

FA 6: (1) F1: Diapason (.80, .98).
(2) F2: Diapason (-.60, .00). Cf. Swahili (-.19).
(3) F3: Diapason (-.30, -.16).
(4) F4: Diapason (-.07, -.05) (intersects with class GH).

CLASS A

FA 4: (1) F2: Diapason (.53, .72); comes close to Yiddish (.52), on the one hand, and Russian (.73) on the other.
(2) F3: Diapason (-.36, -.20). Exception: Mari (.00). Cf. Class I.

Table 7
Factor analysis

	FA 4				FA 6					
	F1	F2	F3	F4	F1	F2	F3	F4	F5	F6
CLASS I										
Vietnamese	.77	-.63	-.12	.00	.82	-.54	.13	-.01	.07	.02
Khmer	.77	-.62	-.12	.00	.84	-.52	.14	-.05	.05	.00
Thai	.79	-.61	-.10	.00	.84	-.54	.10	-.05	.05	.00
O.Chinese	.79	-.61	.00	.00	.84	-.52	.07	-.04	.08	.01
Chinese	.91	-.36	-.17	.00	.92	-.33	.14	.01	.06	.10
Indonesian	.73	-.67	.00	.00	.81	-.57	.08	.01	.00	.05
Tangut	.83	-.51	.19	.08	.88	-.45	-.08	.01	.02	-.01
Tagalog	.79	-.57	.00	.13	.83	-.53	.00	.07	-.07	-.05
Maninka	.79	-.61	.00	.00	.84	-.54	-.02	-.04	.02	-.02
Tibetan	.82	-.44	-.19	-.28	.86	-.41	.16	-.05	-.03	.26
Burmese	.95	-.25	-.16	.00	.97	-.16	.16	-.10	-.01	-.04
Persian	.99	.00	-.09	.00	.98	-.02	.03	.10	.05	-.01
Tadzhik	.99	.00	-.09	.00	.97	-.07	.00	.10	.05	.04
English	.88	-.32	.34	-.08	.89	-.31	-.28	-.15	-.05	-.02
CLASS A										
Mongolian	.84	.39	-.31	-.20	.77	.50	.31	-.21	.02	.04
Korean	.75	.53	-.37	.00	.63	.66	.33	-.18	.03	-.16
Turkish	.79	.54	-.27	.00	.62	.71	.22	-.11	.20	-.08
Mari	.82	.55	.00	-.12	.66	.72	.02	-.08	.04	-.04
Japanese	.57	.72	-.36	-.09	.53	.77	.31	-.12	-.02	-.02
Telugu	.54	.72	-.26	-.33	.47	.80	.25	-.21	-.01	.15
Manchur.	.94	.22	-.23	-.10	.86	.38	.25	-.19	.06	-.04
Chukchee	.70	.63	-.26	.18	.65	.67	.19	.24	.18	-.02
CLASS F										
Sanskrit	.29	.91	.20	-.19	.31	.90	-.14	.01	-.11	.25
Russian	.47	.73	.46	-.10	.47	.75	-.40	.00	-.08	.17
CLASS GH										
German	.78	.20	.59	.00	.71	.24	-.63	-.10	.08	-.13
Hindi	.82	.36	.45	.00	.68	.45	-.53	-.16	.09	-.16
Urdu	.84	.32	.43	.00	.78	.37	-.48	-.14	.05	-.05
CLASS-FREE LANGUAGES										
French	.85	.00	.50	.18	.86	-.09	-.40	.15	-.21	-.07
Yiddish	.65	.52	.46	.29	.75	.21	-.38	.40	-.15	.18
Arabic	.34	.82	-.12	.40	.39	.80	.07	.40	-.07	-.11
Swahili	.33	.19	-.38	.82	.51	.19	.32	.74	-.01	-.11

FA 6:
(1) F1: Diapason (.53, .63). Exceptions: Manchurian (.86), Mongolian (.77), Telugu (.47).
(2) F2: Diapason (.38, .80) (intersects with class F).
(3) F3: Diapason (.19, .33). Exception: Mari (.02). Cf. Class I, Swahili.
(4) F4: Diapason (-.21, -.08) (intersects with class GH). Exception: Chukchee (.24).

CLASS F

FA 4:
(1) F2: Diapason (.73, .91). Includes Arabic (.82).
(2) F3: Diapason (.20, .46). Adjacent to class GH (.43, .59).

FA 6:
(1) F1: Diapason (.31, .47). Includes Arabic (.39). Cf. Telugu (.47)
(2) F2: Diapason (.75, .90) (intersects with class A). Includes Arabic (.80).
(3) F4: Value .00 (intersects with class I).
(4) F5: Diapason (-.11, -.08). Cf. Arabic (-.07), Tagalog (-.07), English (-.05).
(5) F6: Diapason (.17, .25). Cf. Telugu (.15), Tibetan (.26).

CLASS GH

FA 4:
(1) F2: Diapason (.20, .36). Cf. Swahili (.19), Mongolian (.39), Manchurian (.22).
(2) F3: Diapason (.43, .59). Cf. Yiddish (.46), French (.50), Russian (.46).

FA 6:
(1) F1: Diapason (.68, .78). Cf. Yiddish (.75), Mongolian (.77).
(2) F2: Diapason (.24, .37) (intersects with class A). Cf. Yiddish (.21).
(3) F3: Diapason (-.63, -.48).
(4) F4: Diapason (-.16, -.10).

The following comments may be made concerning the above-given criteria:

(a) FA 6 is more effective in discriminating the language classes than FA 4.

(b) Correlational analysis (with its various generalized coefficients) is more effective as a classificatory procedure than factor analysis. The advantage of the latter is in its more differentiated approach, giving the investigator greater opportunity to establish finer distinctions and relations; but it is ill-adapted to

serve as a primary classificatory criterion because of the relatively high degree of equivocality of its results.

Another conclusion to be noted is the high degree of typological diversity characteristic of the Indo-European languages studied in this paper; three of these belong to different typological classes while the fourth - French - is qualified as a class-free language, though closely connected with English. The singling out of German (together with Hindi and Urdu) into a separate typological class distinct from the standard classes of isolating, agglutinative and inflective languages requires further consideration, especially in the light of its "nuclear" position in the class-system, manifested by the highest value of its general average extraclusteral coefficient (.58).

A comparison of the above-given classification of languages with that in Altmann & Lehfeldt (1973) may be effected only within the limits of the six languages common to both lists, i.e. Vietnamese, Persian, English, Sanskrit, Turkish, Swahili. In both classifications Sanskrit and Turkish pertain to different classes, Swahili is an isolated language, while Persian and English are referred to the same class. The only distinction is the separation of Vietnamese from Persian/English in Altmann & Lehfeldt into a separate class. But this distinction, determined as it is by the class-limiting criteria introduced by the investigator, is of a relative nature: the reader may be reminded that in our scheme Vietnamese is referred to the nuclear, Persian and English - to the marginal (i.e. the farthest removed from the nuclear) levels of class I.

These data are insufficient for any far-reaching comparisons and conclusions. But the results, limited as they are, seem encouraging enough to suggest the expediency of combining the two lists into a single data base, open to any further extensions, and treating it on the basis of a common set (or sets) of criteria.

3. Selection of diagnostic typological indices

One of the main tasks of any inductive science is the selection of the diagnostic features of the phenomena studied, i.e. such as represent the whole (potentially infinite) set of features in its significant relations with other sets of features. In our case the task is to single out the minimal set of typological indices which is optimally equivalent to the full set in its classificatory function.

To achieve this aim we shall take recourse to a taxonomic procedure elaborated by V.N. Vapnik (1979). A set of elements, each of which is qualified by a set of quantitative characteristics, is structured on the basis of the notion of "distance". An element is provisionally taken as the starting point of the procedure. Each consecutive stage of the procedure consists in finding that element which is cumulatively the closest to the whole set of elements ranged at the preceding stages.

Our approach is to apply this algorithm first to the whole set of 11 typological indices and then to all possible subsets. It was found that only one of the subsets gives a sequence of languages essentially equivalent to the sequence based upon the full set of indices.

This diagnostic subset consists of the indices of **suffixation, agglutination** and **prefixation.**

The corresponding sequences are given in Table 8.

Table 8

	CLASS 1	Typological distance
1.	Vietnamese	0
2.	Maninka	3
3.	Thai	3
4.	Old Chinese	4
5.	Khmer	7
6.	Indonesian	8
7.	Tagalog	8
8.	Tangut	31
9.	Chinese	31
10.	Tibetan	15
11.	Burmese	10
12.	Persian	20
13.	Tadzhik	5
14	English	44
	CLASS 2	
15.	French	45
16.	Yiddish	41
17.	German	22
18.	Hindi	27
19.	Urdu	9

	CLASS 3	
20.	Manchurian	46
21.	Mongolian	21
22.	Korean	8
23.	Turkish	9
24.	Mari	29
25.	Japanese	31
26.	Telugu	11
27.	Chukchee	34
	CLASS 4	
28.	Russian	47
29.	Sanskrit	43
30.	Arabic	49
31.	Swahili	107

The above-given "Vapnik" (V) sequence of languages corresponds closely with the one established in the course of our investigation on correlational criteria. Ascribing a critically delimiting function to distances not lower than 45 we come to the following differentiation of our set of languages:

(1) Class 1, 3 and 4 of the V-scheme correspond exactly with our classes I, A and F. The mutual correspondence of the first two pairs of stratified classes holds as regards not only their content, but likewise their internal differentiation into three hierarchical levels.

(2) Arabic and Swahili figure as class-free languages in both schemes.

(3) The only divergences pertain to our "problematic" class GH and class 2 of the V-scheme which includes, besides German, Hindi and Urdu, two class-free languages of our scheme: French and Yiddish. As we have seen, French performs a "connecting" function (and thus occupies an intermediate position) between classes I and GH. The difference between the two schemes is a "shift of stress": French is more closely affiliated with class GH in the V-scheme and with class I in ours. Yiddish, in its turn, is critically correlated with French. The MEC's of both Yiddish and French are the highest as regards class GH (.74 and .79 respectively).

We see that these divergences cannot be regarded as crucially important, with the conclusion that the V-classification, based upon three indices, is practically

isomorphic with ours. The question as to which of the two schemes is the more adequate one requires further investigation on a wider set of languages.

There are sufficient grounds therefore for regarding suffixation, agglutination and prefixation as the crucial, diagnostic typological indices constituting a minimal set of characteristics sufficient to serve as a basis for a valid classification of the whole range of empirical data under consideration. The following motivation of the diagnostic status of these three indices may be proposed.

As was shown above, suffixation is the dominant element of cluster A encompassing six indices, while agglutination has the same status in cluster B (three indices). These two indices may therefore be considered as representative of 9 indices out of 11. Of the remaining two isolated indices (prefixation and analyticity) the first is more closely integrated into the system, as illustrated by its medial extraclusteral correlations with the two clusters. It follows that these three indices in their conjunction are the most adapted to represent adequately the whole set of indices in its classificatory function.

The following general conclusions may be drawn from the foregoing discussion.

1. The ten typological indices proposed by J.H. Greenberg have proved their diagnostic efficiency in language classification on a sufficiently wide range of typologically heterogeneous languages (encompassing 45 languages of our and the Altmann & Lehfeldt lists, treated on different criteria but with essentially similar results).

The eleventh index - analyticity - does not fulfil any significant classificatory function and may therefore be discarded from further consideration.

2. The three most diagnostic indices are: suffixation, agglutination and prefixation.

3. The classification arrived at in the course of our investigation has shown a relatively high degree of correspondence with the traditional genealogical and typological language classes. The four representatives of the Indo-European languages (English, French, German, Russian) constitute the most salient exception, being dispersed under different classification headings.

4. The main novelty of our classification is the class GH (German, Hindi, Urdu), which occupies the most central position in the system but cannot be unequivocally brought under any of the traditional classificatory categories. The

introduction of new languages into the sphere of investigation and their treatment on the basis of the same set of criteria will show whether we have here a pseudoclass or a new typological category, the content and qualitative definition of which constitute an appealing task for future exploration.

References

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

Cowgill, W. (1963). A search for universals in Indo-European diachronic morphology. In: Greenberg, J.H. (ed.), *Universals of language: 91-113.* New York.

Greenberg, J.H. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics 26/3, 178-194.*

Kasevič, V.V., Jachontov, K.S. (eds.) (1982). *Kvantitativnaja tipologija jazykov Azii i Afriki.* Leningrad, Izd. Leningradskogo Universiteta.

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

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

Pierce, J.E. (1966). Sampling and typological indices in language. *Linguistics 24, 43-50.*

Silnickij, G.G., Jachontov, K.S., Jachontov, S.E. (1986). Primenenie korreljacionnogo i faktornogo analiza v tipologii jazykov. In: Bartkov, B. (ed.), *Aktual'nye voprosy derivatologii i derivatografii: 105-125.* Vladivostok, Dal'nevostočnoe Otdelenie Akademii Nauk SSSR.

Vapnik, V.N. (1979). *Vosstanovlenie zavisimostej po empiričeskim dannym.* Moskva, Nauka.

Glottometrika 14, 1993, 161-189

Über den gegenwärtigen Stand der automatischen Textverarbeitung in der Forschungsgruppe "Sprachstatistik" (Zum Problem des linguistischen Automaten)

W. Czyżakowski, R. Piotrowski

Einleitung

Bereits seit 35 Jahren arbeitet die Forschungsgruppe "Sprachstatistik" (SpSt) im Bereich der automatischen Textverarbeitung (AT) und der linguistischen Aspekte der künstlichen Intelligenz. Ungeachtet des kritischen Zustandes unserer Informatik und automatischen Textverarbeitung [12: 3,5,9-10,17] bestehen in der GUS diejenigen Kollektive fort, welchen es gelungen ist, AT-Systeme aufzubauen und in Betrieb zu nehmen und den Anforderungen der gegenwärtigen Informationstechnologie gerecht zu werden. Zu ihnen gehört die genannte Forschungsgruppe. Nach Erleben der Krise der maschinellen Übersetzung Anfang der 60er Jahre und in der Folgezeit erarbeiteten Strategien, die jeweils der aktuellen Rechnertechnik angepaßt wurden, wandte sich die SpSt-Gruppe Ende 1970 der Schaffung international konkurrenzfähiger polyfunktioneller AT-Systeme zu [31].

Was hat den SpSt-Kollektiven von Sankt-Petersburg, Minsk, Kischinjow und Mittelasien geholfen, das Schicksal Dutzender wissenschaftlicher Gruppen zu vermeiden, die, nach einigen Jahren des Theoretisierens und vergeblicher Versuche, ein lauffähiges System zu erstellen, aus der Szene der Informatik verschwanden?

Man kann zwei hauptsächliche Gründe für die Lebensfähigkeit der SpSt-Gruppe bestimmen:

Ein Grund ist die Einsicht in die prinzipiellen Unterschiede zwischen der Rede- und Denktätigkeit des Menschen und den intellektuell-linguistischen Möglich-

keiten des Computers. Das Begreifen dieser Unterschiede ermöglichte es der SpSt-Gruppe, ihre Anstrengung auf die Erarbeitung des reproduzierenden 'Rede- und Denktätigkeit'-Analogons des Menschen zu konzentrieren, d.h. auf die Schaffung eines solchen Modells, welches nicht nur eine abstrakte Formulierung und mathematische Definition vorsieht, sondern auch ihre Realisierung in der Form eines wirklich arbeitenden linguistischen Automaten [16:11-13; 21:112-119][1].

Ein zweiter Grund ist die flexible Organisation der wissenschaftlich-technischen und Kaderfinanztätigkeit der SpSt-Gruppe, die sich prinzipiell von bürokratischen sowjetischen Forschungs- und Produktionsunterabteilungen der 60er Jahre unterschied.

1. Theoretische Untersuchungen in der SpSt-Forschungsgruppe

Wenden wir uns den theoretischen Grundlagen und der Technologie der Tätigkeit der SpSt-Gruppe zu, so bemerken wir, daß eines ihrer hauptsächlichen Prinzipien das Studium der Erfahrungen (darunter auch der negativen) sowjetischer und ausländischer AT-Kollektive war.

So hat die Analyse des fehlgeschlagenen Versuchs der Pioniere der maschinellen Übersetzung [15] der SpSt-Gruppe geholfen, eine solche wissenschaftliche Strategie und informative Technologie zu erarbeiten, welche um die 80-90er Jahre den Aufbau tatsächlich arbeitender AT-Systeme ermöglichte.

Als erste Lehre aus den Erfahrungen der Pioniere ergab sich, daß die Erstellung von "Papier"-Algorithmen, die ein Feedback in der Form konkreter Implementierungen und Bewertungen ihrer Resultate von seiten der Benutzer nicht vorsahen, zwar eine intellektuell attraktive Spielerei darstellt und manchmal von sprachwissenschaftlichem Interesse war, wenig für den realen Fortschritt der AT ergibt. Ebenfalls wurde es klar, daß die Computerlinguistik zu den Wissenschaften gehört, deren Ergebnis völlig von der Zuverlässigkeit des Feedbacks zwischen Theorie und Experiment abhängt.

[1] Unter einem linguistischen Automaten wird ein Komplex von Hardware und Software verstanden, der in sich vereinigt:
- eine linguistische Informationsdatenbank [2];
- eine linguistische Software;
- eine Benutzeroberfläche, die das Betriebssystem, die Rechnerressourcen sowie linguistische Informationsdatenbank und linguistische Software zugänglich macht.

Die zweite Lehre war, daß effiziente maschinelle Algorithmen nicht auf der Grundlage traditioneller Grammatiken und Wörterbücher aufgebaut werden dürfen, die auf die Beschreibung statischer Sprachsysteme orientiert sind, sondern auf der Basis lexikalisch-grammatischer Beschreibungen lebendiger dynamischer Sprache, genauer derjenigen Gesamtheiten von Texten (Teilsprachen), auf deren Verarbeitung diese Algorithmen orientiert sind.

Drittens besteht die wichtigste Lehre darin, daß der Aufbau real arbeitender AT-Systeme nur unter der Voraussetzung gelingen kann, daß - und sei es auch nur in den allgemeinsten Zügen - die Spezifik der Textverarbeitung im System Mensch-Computer klar wird.

Die SpSt-Gruppe entwickelte ihre theoretischen Untersuchungen in drei Richtungen.
Die erste Richtung, deren Ziel es war, die linguistische Wechselwirkung von Mensch und Computer zu erfassen, deckte die "genetischen" Paradoxa des Menschen und des Roboters auf, die eine unsichtbare trennende Barriere schaffen, welche die natürliche Sprache des Menschen von der künstlichen Sprache des Computers scheiden [6:168-169; 14:234-266; 22:30-48]. Als die wichtigsten unter diesen Antinomien erwiesen sich:

1) die Nichtübereinstimmung zwischen der kontinuierlichen Natur der Sprache, die auf unscharfen toleranten linguistischen Mengen beruht, und der diskreten künstlichen Sprache, die mit scharfen, äquivalenten Mengen auf die Beschreibung im Computer hin orientiert ist;

2) der Widerspruch zwischen dem offenen, dynamischen (diachronischen) Charakter der lebendigen Sprache und ihrem geschlossenen (synchronen) Begriff im Computer;

3) der Widerspruch zwischen dem einen Sinn des (vom Computer) zu verarbeitenden Textes und den verschiedenen Aspekten der von Mensch zu Mensch übermittelten sprachlichen Botschaft; jede Mitteilung kann nämlich drei Sinne enthalten: den die Pragmatik der Kommunikanten bedingten Autorsinn und perzeptiven Sinn sowie den von dieser Pragmatik unabhängigen universalen Sinn [5:27-30].

Es wurde klar, daß die Konstruktion real arbeitender AT-Systeme stark davon abhängt, inwieweit es gelingt, die Wirkung der erwähnten Antinomien zu neutralisieren.

Die zweite Richtung der SpSt-Untersuchungen orientiert sich auf die Suche nach Wegen zur Neutralisierung der aufgezählten "genetischen" Paradoxa. Diese Untersuchungen wurden der informationstheoretisch-statistischen Struktur des Textes gewidmet. Diese auf der Grundlage slawischen, germanischen, romanischen, türkischen und finnisch-ugrischen Sprachmaterials durchgeführten Untersuchungen ergeben folgende für den Aufbau und die Entwicklung von AT-Systemen wichtigen informationstheoretischen Eigenschaften des Textes.

Alle untersuchten Sprachen, unabhängig von ihrem Typ, besitzen eine im Intervall 65-96% liegende Redundanz. Die größte Redundanz und daher die größte Explizitheit und Affinität zur Computersprache zeigen geschäftliche und Patentdokumente (85-96%), dann folgen wissenschaftlich-technische und publizistische Texte (ungefähr 80%). Am wenigsten redundant erwiesen sich die schöngeistigen Texte und nichtnormierte mündliche Konversation [22:155-160]. Gleichzeitig erwies es sich, daß es dem Rezipienten im Allgemeinen zum Textverständnis genügt, aus dem Text über die Lexik etwa 70% Information herauszuziehen [19:220].

18% (in agglutinierend-synthetischen Sprachen) bis 35% (in der analytischen englischen Sprache) der syntaktischen und semantischen Information ist in Kontextverbindungen des Textes konzentriert. Dementsprechend ist zwischen 65% und 82% der Information im Lexikon des Textes angelegt. Hier hat es sich erwiesen, daß die kontextlose Auffindung dieser Wortformen eine U-förmige Verteilung hat: Die hauptsächliche Masse der Information fällt auf den "lexikalischen" Anfang und das "morphologische" Ende der Wortform, aber ihre Mitte erweist sich meist überflüssig. Diese Informationsverteilungen nehmen innerhalb des Kontexts L-Form an. Der Übergang der U-förmigen in L-förmige Konfigurationen kommt daher, daß der lexikalisch-grammatische Kontext die Informationsmaxima am Wortende wegschneidet. Eben diese Maxima konzentrieren in sich, wie schon gezeigt, den Hauptteil der im Wort enthaltenen grammatischen Charakteristika.

Aufgrund aller dieser Daten wurde die Annahme aufgestellt, daß die maschinelle Verarbeitung allein lexikalischer Einheiten, d.h. von Wortformen und Wortverbindungen ohne syntaktische Analyse des Satzes manchmal dem Text eine solche Menge von Information entnehmen kann, welche ausreichend zum Verständnis seines allgemeinen Inhaltes ist. Diese Annahme betrifft in erster Linie die synthetischen Sprachen, in welchen Wortformen im Durchschnitt um 20-30% mehr Information enthalten als in der analytischen englischen Sprache [38:243-245]. Gleichzeitig haben Informationen über die U- und L-förmigen Informationsstrukturen die Voraussetzung für eine komprimierte Codierung der Text- und Wortinformation im Computer geschaffen.

Schließlich veranlaßte uns die morphologische Redundanz in Textwortgebräuchen, nichttriviale Lösungen bei der Projektierung der Algorithmen zur grammatischen Analyse zu suchen.

Statistische Untersuchungen wurden hauptsächlich zur Lexik, Phraseologie und Textmorphologie der erwähnten Sprachen und in geringem Maße in bezug auf die Syntax vorgenommen. Diese Untersuchungen erbrachten folgende wichtigen Resultate für die Computerlinguistik:

1. Während die Hilfswörter und die häufigsten Vollwörter relativ stabile Häufigkeiten auch in Texten haben, die zeitlich ziemlich weit voneinander entfernt sind und auch verschiedenen Stilen und Untersprachen angehören, hängt die Gebräuchlichkeit terminologischer Wörter stark von der Entstehungszeit des Textes und seiner Zugehörigkeit zu einer bestimmten Teilsprache ab. Daraus folgt, daß man beim Aufbau des automatischen Wörterbuchs die Hilfs- und die allgemein gebräuchliche Lexik aus einsprachigen erklärenden und Häufigkeitswörterbüchern entnehmen darf. Hingegen ist die terminologische Lexik besonderen Gesetzmäßigkeiten schneller Erneuerung unterworfen (23:55-59, 64-68; 32) und muß aus solchen Häufigkeitswörterbüchern entnommen werden, welche mittels Verarbeitung von Gegenwartstexten der betreffenden Fachsprache erstellt wurden. Die auf dieser Grundlage aufgebauten automatischen Wörterbücher müssen periodisch hinsichtlich neuer Termini aktualisiert werden, die aus den im Computer verarbeiteten Texten entnommen wurden.

2. In den heutigen Sprachen wird die Mehrzahl der für das Verstehen von Fachtexten wichtigen Begriffe mit Hilfe von Wortverbindungen ausgedrückt. Deshalb stehen neben traditionellen Häufigkeitswörterbüchern der Wörter und Wortformen Häufigkeitslisten der Wortverbindungen, die auf konkrete Fachsprachen orientiert sind. Auf der Grundlage dieser Häufigkeitsliste müssen große phraseologische automatische Wörterbücher erstellt werden. Die Erfahrung der SpSt-Forschungsgruppe zeigt, daß die Anwendung maschineller Wörterbücher phraseologisch (besonders komplexer Termini) die Qualität der AT isolierender (Chinesisch) und flektierend-isolierender (Englisch, Französisch) Sprachen beträchtlich erhöht.

3. Mehrjährige Vergleiche der relativen Häufigkeit des Erscheinens der lexikalischen Einheit in separaten Abschnitten des Texts mit theoretischen Verteilungen (normale, lognormale, Poisson-, Cebanov-Fucks-, Markov-Kolmogorov- u.a.) (8:118-131; 9:345-360; 10:11-14; 23:353-357) ergaben, daß die Übereinstimmung/Nichtübereinstimmung empirischer Verteilungen der Wortformen und Wortverbindungen mit bestimmten theoretischen Modellen

als formales Merkmal ihrer Zugehörigkeit zu einer bestimmten lexikalisch-grammatischen Klasse dienen kann. So gehorchen z.B. die Artikel, Hilfswörter, Adverbien und Numeralia gewöhnlich binomialen, normalen oder lognormalen Gesetzen und auch den Gesetzen von Poisson und Čebanov-Fucks, Verben unterwerfen sich meist den binomialen und Poisson-Gesetzen, während terminologische Substantive überhaupt keinem der betrachteten theoretischen Modelle gehorchen.[2] Diese statistischen Eigenschaften der lexikalischen Einheiten können bei der Konstruktion des Algorithmus der formalen morphologischen Analyse der Wortformen und Wortverbindungen im Text ohne Benutzung des automatischen Wörterbuches verwendet werden.

Die dritte Richtung wurde Untersuchungen auf dem Gebiet der Generierung und der Rezeption des Textes gewidmet. Trotz des Nutzens informationstheoretisch-statistischer und traditioneller linguistischer Untersuchungen konnten diese nicht das Fundament bilden, auf dessen Grundlage sich die Architektur eines Computer-Analogons der Rede- und Denktätigkeit des Menschen schaffen ließe. Eine solche theoretische Basis muß in semiologischen Hypothesen der zwischenmenschlichen Kommunikation und Mensch-Maschine-Kommunikation gesucht werden. Dieser Aufgabe waren die Ende der 60er Jahre begonnenen Untersuchungen auf dem Gebiet der maschinellen Semiotik gewidmet, und ebenfalls die Suche nach Modellen, die die Generierung und die Rezeption von Texten beschreiben und auf deren Grundlage man die für die Mensch-Maschine-Kommunikation adäquaten maschinellen Prozeduren der Analyse und Synthese des Textes aufbauen könnte.

Einerseits ging hier die Suche in Richtung der Verarbeitung schon bekannter linguistischer, psycholinguistischer und kognitiver Modelle und ihrer Adaptation für die Bedürfnisse der Mensch-Maschine-Kommunikation [1:21-49; 4:221-237; 13:187-200,217,250; 33:124-128,131-132], und andererseits erfolgte sie vermittels der Organisation von zur ökologischen Betrachtungsweise orientierten und von maschineller Metaphorik freien selbständigen Untersuchungen auf dem Gebiet der psychiatrischen Linguistik.

Als Ergebnis dieser Suchaktion wurde ein erweitertes Saussuresches Modell des sprachlichen Zeichens aufgebaut [16:27-47; 28:6-29], auf dessen Grundlage das semiotische kommunikative Schema entwickelt wurde, welches verschiedene

[2] Anm. d. Hrsg.: Diese Zusammenhänge gehorchen sämtlich dem sog. Frumkina-Gesetz, vgl. Altmann, G., Burdinski, V., Towards a law of word repetitions in text-blocks. Glottometrika 4, 1982, 147-167.

Hypothesen der stufenweisen Generierung und des stufenweisen Empfangs einer Nachricht inkorporiert.

Dieses Schema (Abb.1), welches die psycholinguistische Grundlage zum Aufbau "intellektueller" AT-Systeme ist, beschreibt die Bildung der Nachricht, angefangen von ihrer denotativen Idee (Dn), die ein Faktum der objektiven Realität widerspiegelt, über den thematisch-rhematischen designativen Plan zu ihrer lexikalisch-grammatischen und graphemisch-phonetischen Kodierung. Diese Entwicklung steuert der kommunikativ-pragmatischen Operator [36:108], welcher bei der Formierung der Nachricht den Übergang von einer Ebene zur nächsten regelt, indem er die Auswahl der nötigen Information aus dem Thesaurus (Θ) und der linguistischen Kompetenz leistet.

Was den Empfang und die Dechiffrierung der Nachricht betrifft, so orientiert sich die SpSt-Gruppe in ihren Untersuchungen an zwei Schemata (Abb.2). Gemäß dem ersten hypothetischen Schema werden die vom Rezipienten empfangenen Schall- oder visuellen (graphischen) Signale mit in der linguistischen Kompetenz gespeicherten sensorischen (phonemisch-phonetischen oder graphemischen) Mustern verglichen. Wenn ein solcher Vergleich ein positives Resultat ergibt, dann wird eine lexikalisch-grammatische Oberflächen-Analyse angeschlossen, einschließlich der Wortverbindungen und Wörter. Danach wird auf der designativen Ebene eine thematisch-rhematische Tiefen-Analyse erzeugt, die sich auf syntaktisch-semantische Information stützt, welche aus der enzyklopädischen linguistischen Kompetenz und der Analyse des Kontextes extrahiert ist. Auf der abschließenden denotativen Ebene schließlich erfolgt eine verallgemeinerte Interpretation der Information, die aus den Nachrichten vorangegangener Ebenen extrahiert wurde. Diese vom Rezipienten ausgeführten Operationen, die von seiner persönlichen Pragmatik, seinen Präsuppositionen und seiner vorherigen Bekanntschaft mit der Situation abhängen, sollen den Rezipienten zur Formierung des Denotats (Dn) führen, d.h., des verallgemeinerten simultanen Bildes des Faktums, von welchem die aufgenommene Mitteilung handelt.

Gleichheit $Dn_1=Dn_2$ weist darauf hin, daß die Nachricht in genauer Übereinstimmung mit der anfänglichen Idee des Senders vom Rezipienten verstanden worden ist. Bei $Dn_1 \neq Dn_2$ ist die Dechiffrierung der Mitteilung der ursprünglichen Idee nicht adäquat.

Gemäß der zweiten Hypothese erfolgt die Suche nach Dn schon im Rahmen der sensorischen und lexiko-grammatischen Dekodierung der Nachricht (Blöcke 4 und 3' in Abb.2). In dieser Anfangsetappe erfolgt die Aufdeckung der Schlüsselrichtpunkte aller Aussagen (einzelner Wörter, Wortverbindungen, ein-

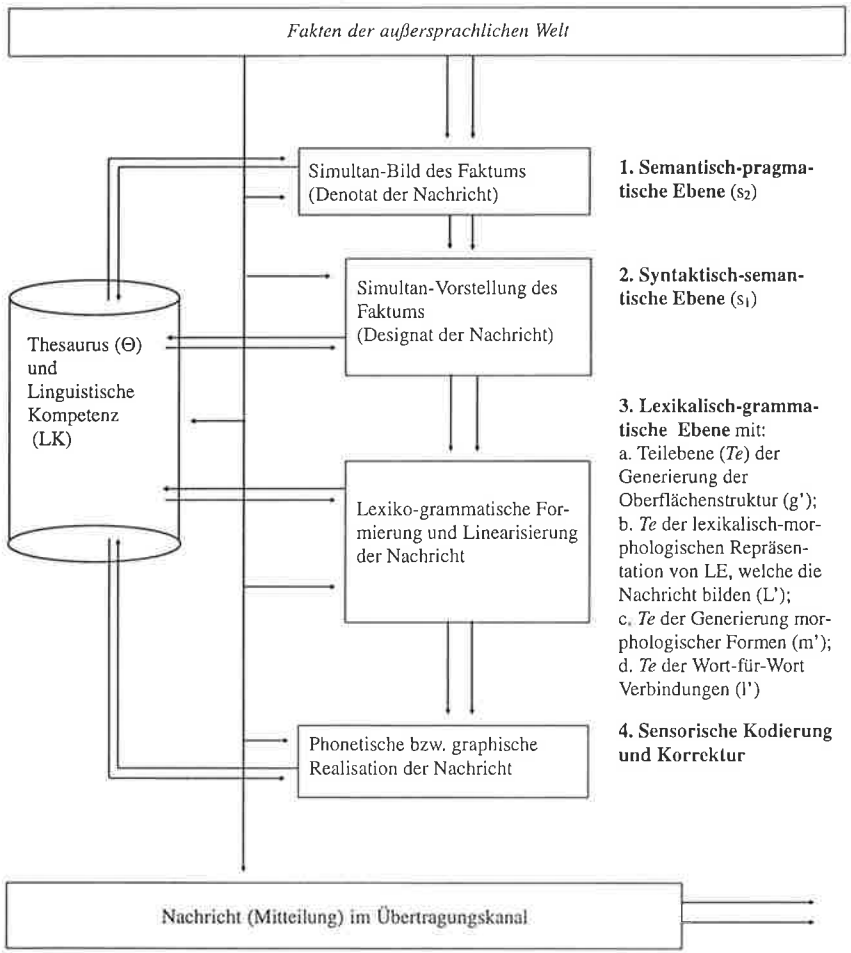


Abb. 1 Darstellung der stufenweisen Generierung (Synthese) einer Nachricht. Strata (Ebenen und Teilebenen) der Modellierung dieses Generierungsprozesses: jedem Stratum entspricht ein Modul des idealen LA. Neben den (in Klammern angegebenen) Kennzeichnungen der Moduln bezeichnen die verwendeten Symbole ⇄ den Prozeß der Spiegelung eines außerweltlichen Faktums im Bewußtsein des Sprechers, ⇓ den Prozeß der Formierung einer Nachricht und ⟷ den kommunikativ-pragmatischen Operator (KPO), welcher die Formierung der Nachricht und die Extraktion von Informationen aus dem enzyklopädischen Thesaurus und der linguistischen Kompetenz steuert.

LE - lexikalische Einheit / WG - Wortgruppe / WV - Wortverbindung

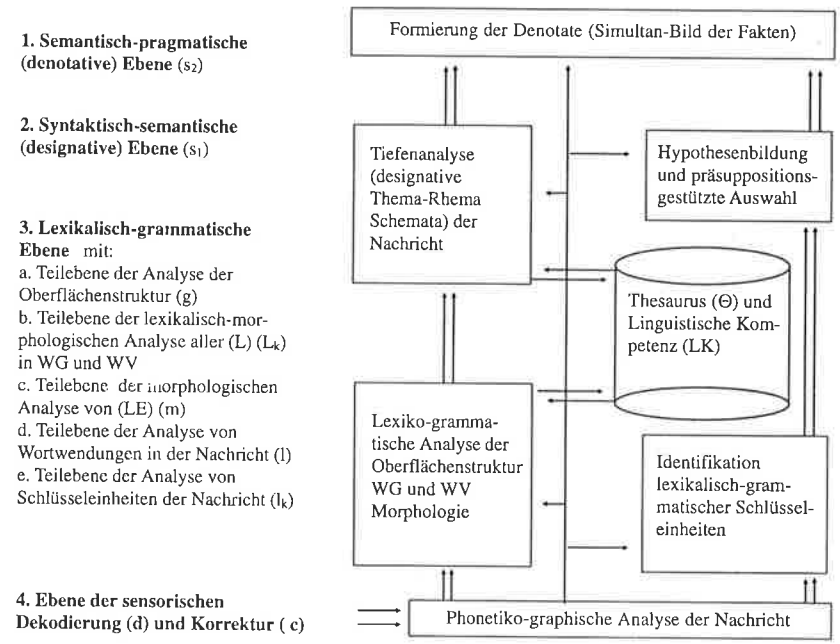


Abb. 2 Darstellung der stufenweisen Analyse (Empfang) einer Nachricht. Strata (Ebenen und Teilebenen) der Modellierung dieses Analyseprozesses: jedem Stratum entspricht ein Modul des idealen LA. Neben den (in Klammern angegebenen) Kennzeichnungen der Modulm bezeichnen die verwendeten Symbole ⇄ den Prozeß der Wechselwirkungen eines außerweltlichen Faktums im Bewußtsein des Sprechers, ⇑ den Prozeß der Rekonstruktion einer Nachricht und ⟷ den kommunikativ-pragmatischen Operator (KPO), welcher die Analyse der Nachricht und die Extraktion von Informationen mithilfe des enzyklopädischen Thesaurus und der linguistischen Kompetenz steuert.

facher semantiko-syntaktischer Schemata). Diese Suche wird vom Rezipienten auf der Basis seiner pragmatischen Einstellung und Erwartung, einschließlich der Orientierung im referentiellen Umfeld und seiner Präsuppositionen durchgeführt. Danach bildet man auf der Grundlage der in den Blöcken 4 und 3' gewonnenen Information eine Hypothese über den Sinn der empfangenen Nachricht. Danach erfolgt aufgrund der pragmatischen Einstellung des Rezipienten, seiner Präsuppositionen wenn, unter Hinzuziehung der lexikalisch-grammatischen Analyse und zuletzt durch Vergleich der erhaltenen Information mit semantisch-syntaktischen Frames, die im Thesaurus und in linguistischer Kompetenz spezifiziert sind, die Wahl der glaubwürdigsten Hypothese, die sich auf ein sinnvolles Bild-Denotat der Nachricht gründet. Die angeführten Schemata werden als Naturobjekte (Originale) für Computermodelle der Analyse und Synthese des Textes betrachtet werden.

2. Der linguistische Automat (LA)

Gestützt auf die semiotische Konzeption der stufenweisen Generierung und des stufenweisen Empfangs des Textes sowie auf den informationell-statistischen Begriff der Textorganisation, begann die SpSt-Gruppe am Ende der 60er Jahre mit der Ausarbeitung einer eigenen Konzeption der AT. Ihre Grundlage bildete die Idee des linguistischen Automaten. Gegenwärtig setzt die SpSt-Arbeitsgruppe die Realisation dieser Idee fort.

Die informationelle und sozio-ökonomische Situation in der zweiten Hälfte des 20. Jahrhunderts bedingte folgende Anforderungen, welchen der LA genügen soll:

1) Polyfunktionalität, d.h. die Fähigkeit des LA, in Abhängigkeit von Bedürfnissen des Benutzers verschiedene Verarbeitungsarten großer Textinformationsströme zu realisieren (Indexierung und Referierung von Dokumenten, ihre maschinelle Übersetzung, Unterstützung des Sprachunterrichts);

2) Minimierung von Informationsverlusten und Überbrückung der Distanz zwischen natürlicher und künstlicher Intelligenz und Sprache;

3) Lebensfähigkeit, d.h. die synergetische Fähigkeit des LA, seine wesentlichen Eigenschaften zu erhalten unter Einwirkung fataler Faktoren wie Ausfall äußerer Anlagen oder Abschnitte des Kernspeichers, Entstellung einiger lexikalische Einheit usw. (Diese Eigenschaft des LA ist sehr aktuell für die wenig zuverlässige russische Computertechnik);

4) Möglichkeit ständiger Entwicklung und Vervollkommnung, welche von der Notwendigkeit der Adaptation des LA nicht nur an die kommunikativ-informationelle Evolution der Gesellschaft, sondern auch an die Pragmatik realer Informationsbenutzer diktiert wird.

2.1 Linguistische Strategie beim Aufbau von LA

Die Erarbeitung der Strategie des Aufbaues von LA verlangte die Entscheidung zweier Alternativen.

Die erste betrifft die Entscheidung zwischen entweder lexikalischer oder grammatischer Priorität beim Aufbau des allgemeinen Algorithmus des LA. Bei der Lösung dieser Aufgabe läßt sich die SpSt-Gruppe von zwei Annahmen leiten:

1) Wie die informativen Untersuchungen des Textes gezeigt haben, trägt die Lexik den Löwenanteil der im Text enthaltenen Information;

2) Maschinelle Analyse und Synthese einzelner lexikalischer Einheiten sind in weit kleinerem Maße als Parsing des Eingangssatzes und Generierung der syntaktischen Struktur des Ausgangssatzes der Wirkung der "genetischen" Paradoxa der AT unterworfen.

Deshalb wurde entschieden, den Aufbau von LA nicht mit der Konstruktion der grammatischer Algorithmen zu beginnen, wie es die Mehrheit der Neophyten der AT taten, sondern mit dem Aufstellen der Wortbasis des LA und der Prozedur der Sinnverarbeitung der lexikalische Einheit des Textes.

Die zweite Entscheidung war mit der Wahl zwischen Chomskyanischen, strikt deduktiven Traditionen und probabilistisch-funktioneller Grammatik der Sprache verbunden.

In gleicher Weise, wie in der Linguistik traditionell lexikalisch-grammatische Modelle zunächst mit Hilfe scharfer Mengen erstellt wurden und dann durch die realistischeren unscharfen Mengen [38:16-17] ersetzt wurden, sollte jetzt bei der Modellierung von Textanalyse- und -syntheseprozessen vorgegangen werden. D.h., es sollten quantitative Bewertungen von typischen Situationen (Frames), Wahrscheinlichkeitsmodelle zur Vereindeutigung und zur semantischen Analyse des Textes herangezogen werden [1:49-54,116-162; 192-236; 29:94-114].

Folglich unterscheidet sich die linguistische Strategie der SpSt-Gruppe von den meisten anderen Kollektiven durch die Priorität der lexikalischen Verarbeitung,

die an die Aufgaben der Benutzer adaptiert ist, zusammen mit der Orientierung auf probabilistisch-funktionelle Linguistik.

2.2 Architektur des LA

Es gibt zwei alternative Betrachtungsweisen hinsichtlich des Aufbaus des LA:

a) seine deduktiv-systematische Entfaltung nach einem starren, von vornherein vorgegebenen Schema "Top-Down" aus dem semantisch-pragmatischen Block zur lexikalisch-grammatischen und Kodierungsebene (Abb.2) und

b) einen iterativen Aufbau des Automaten "Bottom-Up" - von elementaren Wortblöcken zu komplexen lexikalisch-morphologischen, semantisch-syntaktischen und pragmatischen Blöcken.

Trotz ihrer Anziehungskraft hat die erste Betrachtungsweise zwei prinzipielle Nachteile. Erstens erlaubt sie es nicht, gleichzeitig den ganzen Kreis der Aufgaben zu erfassen, die beim Aufbau eines polyfunktionalen LA entstehen. Zweitens gestattet es diese Betrachtungsweise nicht, diejenigen wissenschaftlichen Fortschritte zu nutzen, welche im Laufe der Arbeiten über LA und nach ihrer Beendigung stattfinden, ohne seine Architektur umzugestalten. Insofern erweist sich der nach einem starren deduktiven Schema aufgebaute LA als unfähig, der Wirkung des Paradoxons Mensch-Computer und der Antinomie "Diachronie/Synchronie" zu begegnen (s. oben).

Als konstruktiver erweist sich in dieser Situation die zweite, iterative Betrachtungsweise, welche nicht so sehr von der aktuellen schwierigen Lage unserer Informationstechnologie und Rechnertechnik diktiert ist [12:17] wie von der Notwendigkeit, die oben beschriebenen Antinomien und den von ihnen hervorgerufenen Effekt der Abtrennung maximal abzuschwächen. Diese iterative Betrachtungsweise wird zur Zeit realisiert.

- erstens in ihrer offenen, stufenweisen Organisation, welche einerseits die Möglichkeit in Betracht zieht, aus LA einige Module zu entfernen und andere zu inkorporieren, andererseits das Wechselverhältnis jedes Moduls mit einer bestimmten Ebene der Generierung und des Empfangs der Nachricht zu sichern (Abb.1-2);

- zweitens im Mensch-Maschine-Prinzip seines Funktionierens und seiner Vervollkommnung (s. unten).

LA ist ein komplexes System. Deshalb muß man zu seiner Beschreibung einen vielschichtigen Begriff anwenden, der Modelle und Schemata enthält, die auf hardware, software, lingware und anderen Gesichtspunkten aufgebaut sind. Für uns sind zwei Schemata der Beschreibung besonders interessant: das strukturell-funktionale Schema und das dezisive Schema.

2.2.1 Strukturell-funktionale Beschreibung des LA

Diese Beschreibung, die vom physikalischen Substrat des Automaten abstrahiert, stellt einen Automaten in Form eines hierarchischen Systems dar, welches folgende drei Ebenen (Strata) (24:13) hat:

1. Ein oberes Stratum, welches sich gegenwärtig in bezug auf den LA in Form einer Mensch-Computer-Wechselwirkung realisiert. Diese Wechselwirkung kann man bedingt als Analogon des Motivs und teilweise des kommunikativ-pragmatischen Operators im Schema der Rede-Denktätigkeit des Menschen betrachten (Abb.1 und 2).

2. Das Mittelstratum kann in Form einer Menge von Untersystemen des AT

$F = \{y, Я, Ф, И, А, Э, П, П_3\}$

dargestellt werden, wobei:

y ist ein Untersystem, das die Einheiten des Textes (Buchstaben, Buchstabenkombinationen, Wortformen, Wortverbindungen) nach Alphabet, Frequenz, Alphabet endlicher Buchstaben usw. verwaltet;

Я ist ein Untersystem, das die Sprachzugehörigkeit des untersuchten Textes feststellt;
Ф ist ein Untersystem, das die Fragmentation des Textes leistet;
И ist ein Untersystem, das die Indexierung des Textes und seiner Fragmente durchführt;

A ist ein Untersystem, das die inhaltliche Repräsentation des Dokuments zusammenstellt und seine Referierung durchführt;
ɔ ist das interaktive Experten-Untersystem;
Π ist das Untersystem der maschinellen Übersetzung;
Π_3 ist das Untersystem der thematisch-rhematischen maschinellen Übersetzung der Titel von Artikeln und Büchern[3]

3. Das untere Stratum beschreibt man durch eine Menge (M) funktionaler Module, die aus zwei Untermengen M_a und M_s bestehen. Die erste von ihnen umfaßt die analysierenden Module

M_a : {d, c, l_k, l, m, L_k, L, g, s_1, s_2}

wobei

d = Modul zur Dekodierung des Textes;
c = Modul zur Korrektur;
l_k = Modul zur lexikalischen Analyse der Schlüssel - lexikalischen Einheit des Textes;
l = Modul zur Wort-für-Wort- und Wortverbindung-für-Wortverbindung- (lexikalischen)Analyse aller lexikalische Einheiten des Textes;
m = Modul zur autonomen morphologischen Analyse der Wörter des Textes;
L_k = Modul zur lexikalisch-morphologischen Analyse der Schlüssel-lexikalischen Einheiten des Textes;
L = Modul zur lexikalisch-morphologischen Analyse aller lexikalischen Einheiten des Textes;
g = Modul zur Analyse der Oberflächenstruktur des Textes;
s_1 = Modul zur Analyse der Tiefenstruktur des Textes (Thema-Rhema);
s_2 = Modul der semantisch-pragmatischen Analyse des Textes.

Die zweite Untermenge enthält die Synthesemodule

M_s = {k, c, l', L', g', s_1', s_2'}, wobei

[3] Man kann selbstverständlich die Zahl der funktionellen Untersysteme erweitern. Zum Beispiel kann man experimentelle Module zur Text-Generierung (unter anderem auch von Verstext) aus einem gegebenen sinninhaltlichen Motiv konstruieren. Diese Idee benützt man auch bei der Generierung des Ausgangstexts im Lauf der maschinellen Übersetzung [7]. Es werden auch Versuche gemacht, den LA zum Sprachunterricht zu nutzen. Indem wir das System mit Callware ausrüsten, verwandeln wir es in einen Sprachunterrichtsautomaten [20; 36:106-110;39:1-8].

k = Modul zur graphischen und phonetischen Kodierung des Textes;
c = Modul zur Korrektur;

l' = Modul zur lexikalischen Synthese (Auswahl lexikalischer Äquivalente für Eingangswortformen und -Wortverbindungen aus dem automatischen Wörterbuch);
L' = Modul zur lexikalisch-morphologischen Synthese;
g' = Modul zur Synthese der Oberflächenstruktur des Zieltextes;
s_1' = Modul zur Synthese der Thema-Rhema-Struktur des Zieltextes;
s_2' = Modul zur Synthese der semantisch-pragmatischen Bildes des Textes.

Indem die erwähnten Module mit bestimmten Ebenen und Unterebenen der Generierung und Analyse der Nachricht korreliert sind (Abb.1-2), treten sie als ihre Maschinenanaloga auf. Beim Vergleich verschiedener AT-Systeme ist es zweckmäßig, den Begriff "Arbeitsraum des LA" zu verwenden [25:9]. Für einen idealen LA wird er das Aussehen des kartesischen Produkts S=FxM haben, welches alle Paare <f_j, m_j> (f_i = ein bestimmtes Untersystem,m_j=- ein bestimmtes Modul) enthält. Was den Arbeitsraum S* jedes real existierenden LA* betrifft, so enthält er nur diejenigen binären Verhältnisse f_j m_j, welche für den gegebenen LA* begründet sind. Zum Beispiel wird der Arbeitsraum des einfachsten LA*, welcher die maschinelle Übersetzung Wort für Wort durchführt (18), die Form:

S* = { Πd, Πl, Πl', Πk} haben.

Den Arbeitsraum eines komplizierten LA*, welcher die Fragmentierung, Zusammenstellung der inhaltlichen Repräsentation und Rohübersetzung eines nach Sachrubriken fragmentierten französischen und englischen Patents ausführt [3:17], kann man in Form folgender Mengen darstellen (Siehe Abb.3):

S* = { Φd, ΦL_k, ΦL', Φk, Ad, Al, Al', Ak, Πd, ΠL, ΠL', Πk}.

Französisches Patent

[19] république française
[11] 2.071.055
[21] 69.43575
[15] brevet d'invention
[22] 16 décembre 1969, 16.h. 45 mn.

.....

revendication 2.071.055.1. mécanisme de transmission qui comporte un moteur disposé longitudinalement par rapport à l'axe du véhicule. un embrayage. une boîte de vitesse qui comprend au moins un arbre d'entrée et un arbre de sortie parallèles, un mécanisme de différentiel disposé entre le moteur et la boîte de vitesse

.....

Suchbild des Patents (Wort-für-Wort- und Wortverbindung-für-Wortverbindung-Referierung):
Patentformel, Mechanismus, Antarieb, Motor, Achse, Transportmittel, Kupplung, Wechselgetriebe, Ausgleichsgetriebe
Differenzierung nach konzeptuellen Feldern (Frame-Enquête) und übersetzende Quasireferierung:
[19] patentiertes Land: Frankreich
[11] Nummer des Patents: 2.071.055
[21] Registierungsnummer der Anmeldung: 69.43575
[15] Art der Publikation: Patent
[22] Datum der Anmeldungseinreichung: 16.12.1969, 16 u. 45 M.

Patentformel 2.071.055
Name des Erfindungspatentes:
1. Transmissionsmechanismus
Gesamtheit der spezifischen Charakteristika der Erfindung:
welche aus dem Motor besteht, der längsläufig zur Achse des Transportmittels liegt, Kupplung, Wechselgetriebe:
Gesamtheit der spezifischen Charakteristika des Gegenstands der Erfindung:
welche wenigstens Eintrittswelle und Austrittswelle parallel enthält, Mechanismus des Ausgleichsgetriebes liegt zwischem dem Motor und Wechselgetriebe.

Abb. 3 Konzeptuelle Verarbeitung des französischen Patenttexts und seine MÜ [3:16]

2.2.2 Dezisives Schema der LA

Jede AT ist immer mit Operationen der Erkennung verbunden, welche im Zustand der Unbestimmtheit ausgeführt werden. Diese Unbestimmtheit wird in die linguistische Informationsdatenbank und die algorithmischen Blöcke durch eine Menge von Alternativen übergeben, aus welchen der LA die richtige Entscheidung auswählen muß. Deshalb muß die Architektur von LA nicht nur unter strukturell-funktionalem, sondern auch unter dezisivem Gesichtspunkt beschreiben werden. Bei einer solchen Betrachtungsweise ist der LA durch Abbildung der Eingangs-lexikalischen Einheit (des Eingangstextes T) in die Menge der Ausgangs-lexikalischen Einheit (des Zieltextes T^{out}) unter Einwirkung aller Operatoren G gegeben, welche in der Rolle des Analogons des kommunikativ-pragmatischen Operators auftreten (Abb. 1 und 2.). Mit anderen Worten, wir haben:

$$LA^*:T^{in} \xrightarrow{G^i_j} T^{out} \qquad i=1,...,m;\ j=1,...,n \qquad (1),$$

wobei

m = Anzahl der Untersysteme, die im gegebenen LA* verwendet sind,
n = Anzahl der Module, welche im i-ten Untersystem verwendet sind [24:12-13].

Analog zu den Leitungssystemen kann man das dezisive Schema, das durch den Ausdruck (1) beschrieben wird, in Form einer Stufenhierarchie darstellen:

1) Selbstorganisation,
2) Adaptation des LA an die zu verarbeitenden Texte,
3) Wahl der Lösungsmethode für die beschlossene Aufgaben.

Auf der ersten Stufe, gewöhnlich interaktiv, wird die Strategie der Lösung der allgemeinen Aufgabe P erarbeitet. In Übereinstimmung mit dieser Strategie werden Untersysteme und Module erarbeitet und zusammengestellt, welche der Automat für die Lösung dieser Aufgabe benötigt.

Im Anfangsstadium der Lösung befindet sich der LA gewöhnlich im Zustand der Unbestimmtheit, die einerseits durch Polysemie der Wörterbuch-lexikalischen Einheit, Mehrdeutigkeit der im Text enthaltenen morphologischen Formen und syntaktischen Schemata und auch durch Mangel an in linguistische Informationsdatenbank enthaltenem linguistischem und enzyklopädischem Wissen erzeugt sind. Deshalb muß die dezisive Architektur Adaptationsmittel besitzen, welche diese Unbestimmtheit verringern. Zu ihnen gehören zuerst filtrierende

Algorithmen [38:91-145], siehe auch unten, und zweitens Verfahren der Adaptation des LA an die zu verarbeitenden Texte. Zu letzteren gehören in erster Linie die Ergänzung des automatischen Wörterbuches durch geographische Benennungen, Eigennamen und auch fachsprachenspezifische terminologische Wortformen und Wortverbindungen, Schaffung neuer und Änderung bereits arbeitender Algorithmen. Diese Simulation des LA wird realisiert sowohl interaktiv als auch vollautomatisch durch Auswahl der häufigsten Alternativen [22:296-298]. Der Komplex aller dieser Methoden bildet die zweite adaptive Ebene des dezisiven Schemas des LA.

Das wichtigste Problem für die Entwicklung der Konzeption des LA ist die Organisation und Funktion der Mechanismen des dritten Stratums. Betrachten wir dieses Problem im Detail.

2.2.2.1 Auswahl des besten Verfahrens zur Entscheidung

Gegenwärtig werden in der SpSt-Gruppe zur Auffindung optimaler Lösungen eine Reihe von Verfahren erarbeitet, welche die angewandt-linguistischen Einschränkungen in Betracht ziehen, denen der LA unterliegt. Ein Teil von ihnen ist bereits algorithmisiert. Am interessantesten sind zwei dieser Suchmethoden.

Das erste Verfahren besteht in der hierarchischen Organisation der Arbeit von Untersystemen und Modulen, welche mit der Fähigkeit letzterer zu autonomer Funktion verbunden ist. Dieses Verfahren wird durch folgende Regeln realisiert:

- Die höchste Steuerungsebene besteht in der interaktiven Entscheidung durch den Benutzer,
- Untersysteme und Module der obersten Ebenen bedingen die zielgerichtete Arbeit entsprechender Blöcke der tieferen Ebene,
- Wenn sich die vollautomatisch arbeitenden Untersysteme und Module als unfähig erweisen, eine Entscheidung oder einige alternative Entscheidungen zu fällen, dann werden die hier erhaltenen Resultate der Textverarbeitung zur höchsten Ebene der Hierarchie übergeben, um dort die endgültige Entscheidung zu erarbeiten [11:43-45; 31:29-30].

Betrachten wir diese Prozedur am Beispiel des LA* (Abb.4), welcher die thematisch-rhematische Übersetzung deutscher Titel von zum Gegenstandsbereich "Abwasser" gehörenden Artikeln ausführt [24;26]. Die lexikalisch-grammatische Verarbeitung des Titels des Texts beginnt im Block 2, der die untere Ebene des LA* darstellt. Diese Verarbeitung wird mit Hilfe des deutsch-russischen Wörterbuchs und der maschinellen Morphologie ausgeführt, die in

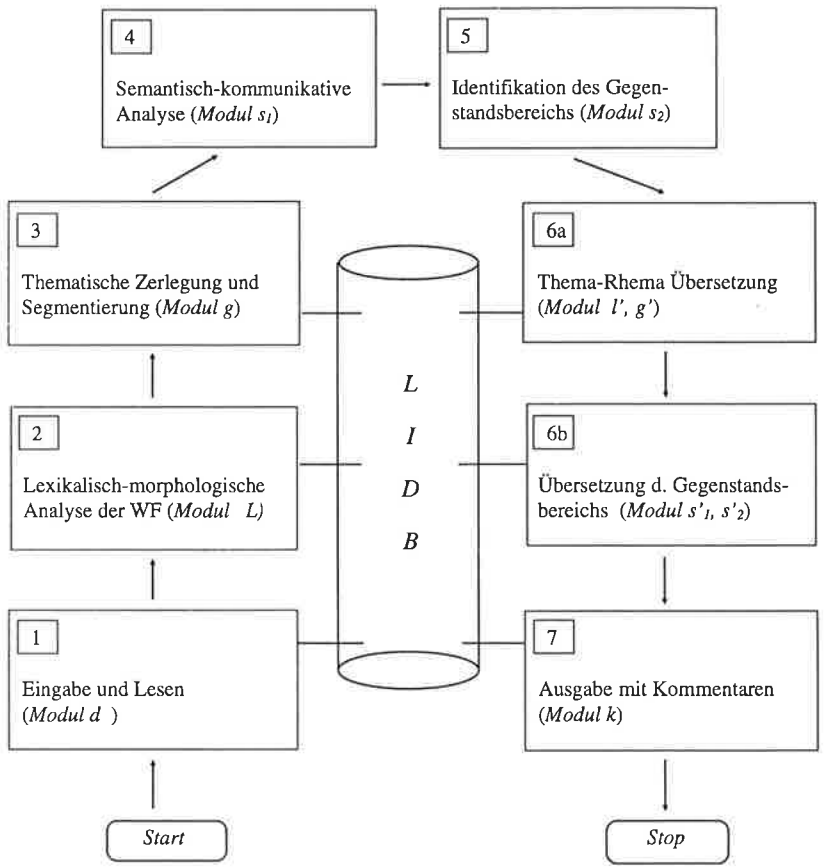


Abb. 4 Strukturell-funktionales Schema der LA* zur Durchführung der thematisch-rhematischen maschinellen Übersetzung deutscher Titelüberschriften

der linguistischen Informationsdatenbank enthalten sind. Sie stellt nicht nur die lexikalisch-morphologische Aufbereitung zur weiteren Übersetzung dar, sondern schließt Information über die morphologischen Grenzsignale und semantische Information ein, die zur Bestimmung der Entscheidungen auf höherer Ebene verwendet werden.

In Block 3 wird mit Hilfe von morphologischen Indikatoren, die im Block 2 erhalten werden, und syntaktischen Grenzsignalen, die aus der linguistischen Informationsdatenbank extrahiert werden, die Aufgliederung des Titels in semantisch-syntaktische Segmente durchgeführt. Zur Bestimmung der kommunikativen (thematischen oder rhematischen) Natur dieser Segmente wird in den Blöcken 3-5 eine Kette von Filtern angewendet, mit deren Hilfe dann eine der alternativen Entscheidungen getroffen wird. Der erste im Block 3 enthaltene Filter ist probabilistisch-syntaktischer Natur. Die vorläufigen statistischen Untersuchungen deutscher Titel zeigten nämlich, daß ungefähr in 90% der Fälle das erste Segment vollständig mit dem Rhema zusammenfällt oder Teil des Rhemas ist. Was die Endsegmente des Titels betrifft, so fallen sie ungefähr in 70% der Fälle in seinen thematischen Teil. Noch weniger genaue Zugehörigkeit zum Rhema oder Thema besitzen das zweite und dritte Segment. Oft liefert also die positionelle Segmentierung des Titels keine eindeutige Entscheidung. Darum muß man die in den Blöcken 2-3 erhaltenen Resultate auf eine höhere Ebene übertragen - in Block 4, wo die weitere kommunikative Analyse des Textes ausgeführt wird und gleichzeitig die Resultate der im Block 3 erhaltenen Segmente kontrolliert werden. Im Block 4 arbeitet der Filter, der die Kontrolle aller Wörter und Stämme des Titels auf ihre Identität mit den lexikalischen Einheit, welche sich in den Listen der in der linguistische Informationsdatenbank enthaltenen rhematischen und thematischen Indikatoren befinden, durchgeführt. Die Resultate dieser Identifizierung werden mit der von der darunter liegenden Ebene (Block 3) übergebenen Informationen verglichen. Diese Operation hat gewöhnlich eines der folgenden Ergebnisse:

1. Die als rhematische Indikatoren bezeichneten lexikalischen Einheiten geraten in die Anfangssegmente, die thematischen Indikatoren dagegen in die Endsegmente. Auf diese Weise fallen die Resultate der Analyse auf beiden Ebenen zusammen, und der LA* nimmt die folgende Entscheidung an: Das Endsegment stellt das Thema dar, das Anfangssegment das Rhema. Die an die thematischen und rhematischen Segmente attributiv angeschlossenen Wortverbindungen können die Determinative des Themas oder Rhemas sein.

2. Die in Block 3 gewonnene Information widerspricht der im Block 4 erarbeiteten Information. In diesem Fall hat die im vierten (obersten) Block erhaltene thematisch-rhematische Segmentierung Priorität.

3. Block 4 liefert durch die thematisch-rhematische Segmentierung des Textes keine eindeutige Entscheidung. In diesem Falle werden die in den Blöcken 3-4 erhaltenen Parameter des Titels an den höchsten 5. Block übergeben, dessen hauptsächliche Aufgabe in einer Zuordnung des Titels zu den Gegenstands-Unterbereichen des untersuchten Gegenstandsbereiches besteht. Zur Entscheidung dieser Aufgabe verwendet man die Indizes der Zugehörigkeit terminologischer Wortformen zum Gegenstand-Unterbereich "Abwasser" und anderen in den untersuchten Texten berührten angrenzenden Gegenstandsbereichen (Diese Indizes sind in den Wörterbuchartikeln zu den entsprechenden Stichwörtern angegeben).

Vorläufige statistische Untersuchungen zeigten, daß man auf die Termini, die zu den Unterbereichen "Abwasser" gehören, am häufigsten im thematischen Teil des Titels trifft, auf die Termini anderer Gegenstandsbereiche dagegen im rhematischen Segment des Titels. Folglich kann man die Anwesenheit im Segment eine Wortform, die zu einem bestimmten Gegenstandsbereich gehört, als ergänzenden Indikator der Rhematizität und Thematizität des Segments verwenden. Illustrieren wir diese Situation am Beispiel der thematisch-rhematischen Analyse des deutschen Titels eines Artikels aus der Zeitschrift "Wasserwirtschaft-Wassertechnik" (1985, Nr. 1. S. 4). Ein Fragment des Ausdrucks seiner Analyse und Übersetzung ist in Abbildung 5 gezeigt. Die Verarbeitung des Titels in den Blöcken 2-4 gibt keine gültige Entscheidung zur Identifizierung von Thema und Rhema. Deshalb werden alle Parameter nach oben an Block 5 übergeben, wo die Zuordnung des zu verarbeitenden Titels zu einem der mentalen Räume (Unterbereiche) des untersuchten Gegenstandsbereich durch Verarbeitung der Indizes der Zugehörigkeit der Termini erfolgt. Dabei betrachtet man diejenigen Termini, welche die Indizes angegebener mentaler Räume besitzen, als schwache Indikatoren des Themas, während die Wortformen und Wortverbindungen mit dem Index irgendeines anderen vom Bereich "Abwasser" verschiedenen Gegenstandsbereiches die Rolle des Exponenten des Rhemas besetzen. Ausgehend von dieser Regel wird z.B. das Segment *Auswerterrechner* in den rhematischen Teil des Titels eingeschlossen. Was das Segment *Trinkwasseraufbereitung* betrifft, so bestätigt sein Gegenstandsindex die Zugehörigkeit dieser Wortform zum thematischen Segment. Die kommunikative Natur des Segments bei Verfahrensuntersuchungen bleibt ungeklärt und muß im Laufe des Kontakts des LA* mit dem Benutzer entschieden werden.

Titel: Einsatz eines Auswerterrechners bei Verfahrensuntersuchungen in der Trinkwassseraufbereitung.*
Einsatz - S, N/D/Ac, Sg, R [941]
eines - Art, G, Sg, Gs [0}
Auswerterrechners - Cp, S, G, Sg, M [105]
Trinkwasseraufbereitung - Cp, S, Cc, Sg, TW [1001]

.....

Einsatz eines Auswerterrechners - Rhema
in der Trinkwasseraufbereitung - Thema
bei Verfahrensuntersuchungen - keine Lösung
Übersetzung: primenenie vyčislitel'noj mašiny pri issledovanii metodov podgotovki pit'evoj vody

Abb. 5 Thematisch-rhematische MÜ des deutschen Titels [26]
Abkürzungen: * - Grenzen der Segmente, Ac - Akkusativ, Art - Artikel, Cc - allgmeiner Kasus, Cp - zusammengesetztes Wort, D - Dativ, G - Genitiv, Gs, - Grenzsignal, M - GB "Maschinen und Ausrüstungen", N - Nominativ, R - Rhema, S - Substantiv, Sg - Singular, TW - GUB "Wasserbewirtschaftung", in Klammern sind die Adressen russischer Äquivalente angegeben.

Das zweite Verfahren zur optimalen Entscheidung besteht in der Fähigkeit des LA* zur Dekomposition oder Vereinfachung der allgemeinen Aufgabe P in dem Falle, wenn ihre Entscheidung unmöglich ist oder ein Zeitaufwand und Ressourchen des Gedächnisses erfordert, über die der LA* im gegenwärtigen Moment nicht verfügt.

Im Falle der Dekomposition wird die allgemeine Aufgabe in Form von Mengen spezieller Aufgaben dargestellt

$P = \{P_1, P_2, ..., P_i, ..., P_n\}$

Als Beispiel betrachten wir die Situation beim Aufbau einer experimentellen türkisch-russischen maschinellen Übersetzung. Der Nichtisomorphismus nominaler und verbaler Flektionsparadigmen in der türkischen und der russischen Sprache ist äußerst groß. Deshalb sind die lexikalisch-morphologischen Module L und L' unfähig, ohne die Hilfe der Module der Analyse und Synthese der Oberflächen- und Tiefenstruktur des Satzes (g/g', s_1/s_1') die russischen Wortformen und Wortverbindungen anzugeben, welche morphologisch genau den türkischen Eingangswortformen entsprächen. Da die Module g/g', s_1/s_1' für die türkisch-russische maschinelle Übersetzung bisher nicht existieren, muß man

die lexikalisch-morphologische Aufgabe L/L' in drei unabhängige Unteraufgaben aufteilen:

P_1 - Analyse der türkischen Wortverbindung; ihr Resultat ist die Aufgliederung in den Stamm (Ausgangsform) und dessen Affixkomponenten (vergl. Modul m),

P_2 - Bestimmung der grammatischen Natur jedes Affixes (Modul m'),

P_3 - Übersetzung des Stammes (Module 1/1').

Die Verwendung der Information, welche man als Resultat der Entscheidung der Unteraufgaben erhält, erlaubt es dem Benutzer, selbst die Übersetzung des türkischen Satzes zu formulieren (Abb. 6).

Als typisches Beispiel der Ersetzung der allgemeinen Aufgabe P durch ihre Vereinfachung P dient der Übergang des LA* zur Wort-für-Wort- und Wortverbindung-für-Wortverbindung-Übersetzung in den Fällen, wo zum Aufbau der Oberflächen- und Tiefenstruktur des Eingangssatzes morphologische und semantisch-syntaktische Ressourcen fehlen. Indem das Verfahren der Dekomposition und Vereinfachung von P es erlaubt, aus den Sackgassensituationen, welche beim Vorsage des Automaten von der gegebenen Form der Textbearbeitung entstehen, herauszukommen, erhöht es merklich die Überlebensfähigkeit des LA*.

edinilen bilgiye göre iş bankaşi, önselikle federal almanyada dört ya da beş şube açmayı planlamıstır.

edin il en	polučat' (Passiv Partizip)
bilgi ye	svedenic (Dativ)
göre	soglasno
iş	trud, delo
banka sı	bank (3. Sg.)
önselikle	v pervuju očered'
federal	federativnyj
almanya da	Germanija (Lok.)
dört	četyre
ya da	ili
beş	pjat'
şube	filial
açma yı	otkpytic (Akkusativ)
planla mıstır	planirovat' (Perfekt)

Post-Redigierung:
soglasno polučennym svedenijam trudovoj bank zaplaniroval otkrytie, v pervuju očered' v FRG, četyrëh ili pjati filialov
'Laut erhaltenen Nachrichten hat die Geschäftsbank die Eröffnung von vier oder fünf Filialen in erster Linie in der BRD geplant.'

Abb. 6 Türkisch-Russische lexikalisch-morphologische MÜ mit Post-Redigierung

3. Schlußfolgerung

Theoretisches Endergebnis mehrjähriger Untersuchungen, welche die SpSt-Gruppe hinsichtlich der linguistischen Aspekte der künstlichen Intelligenz durchführte, war die Konzeption des polyfunktionalen Mehrebenen-LA, welcher die Rede-Denktätigkeit des Menschen modelliert. In den letzten 20 Jahren wurden auf der Grundlage dieser Konzeption einige experimentelle und industrielle LA [6:192-236; 37:120-154] konstruiert. Im Westen sind erst in den letzten Jahren die Idee der Konstruktion polyfunktionaler Systeme und erste experimentelle Realisierungen aufgetreten [30;34;35].

Selbstverständlich kann man die existierenden polyfunktionalen Systeme der AT nur bedingt als vollwertige linguistische Automaten betrachten. Es ist nämlich so, daß die entscheidende Rolle in der Leitung des Feedbacks des LA und der Interaktion seiner Module dem Menschen gehört. Dabei ist auf den höheren Ebenen der strukturell-funktionalen und dezisiven Organisation der Beitrag des Menschen zu dieser Leitung höher als in den niederen primitiven Blöcken des LA. Mann kann erwarten, daß sich die Bemühungen der Forscher in den nächsten Jahren auf die Erweiterung der dezisiven Möglichkeiten des LA konzentrieren werden. Sie werden durch Modellierung einer immer größeren Zahl von Funktionen des kommunikativ-pragmatischen Operators verwirklicht werden, der die Rede- und Denktätigkeit des Menschen leitet [36:108].

Bibliographie

1. Apollonskaja T.A., Glejbman E.V., Manoli I. Porožđajuščie i raspoznajuščie mechanizmy funkcional'noj grammatiki. Kišinev: Stiinca, 1987.

2. Beljaeva L.N. Proektirovanie lingvističeskich informacionnych baz dlja sistem avtomatičeskoj pererabotki teksta. StRAPT, 1988, s. 8-28.

3. Botnaru R.V. Avtomatičeskoe raspoznavanie smysla naučnotechničeskogo teksta na osnove reljatornych frejmov. Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata filologičeskich nauk. Leningrad: LGU, 1985, 16 s.

4. Veličkovskij, B.M. Sovremennaja kognitivnaja psichologija. Moskva: Izdatel'stvo MGU, 1982. 336 s.

5. Gončarenko V.V., Šingareva E.A. Frejmy dlja raspoznavanija smysla teksta. Kišinev: Štiinca, 1984, 198 s.

6. Dvujazyčnoe annotirovanie i referirovanie (Piotrovskij R.G., Beljaeva L.N., Popeskul A.N., Šingareva E.A.). Itogi nauki i techniki. Serija "Informatika". Tom 7. Avtomatizacija indeksirovanija i referirovanija dokumentov. Moskva: VINITI, 1983, s. 165-244.

7. Zubov A.V. Verojatonostno-algoritmičeskaja model' porožđenija teksta (semantiko-sintaksičeskij aspekt). Avtoreferat dissertacii na soiskanie učenoj stepeni doktora filologičeskich nauk. Moskva: Voennyj krasnoznamennyj institut, 1985, 45 s.

8. Kaušanskaja, M.V. Luk"janenkov, K.F. Statističeskij otbor ključevych i terminologičeskich edinic dlja atribucii i referirovanija francuzskogo naučno-techičeskogo teksta. Avtomatičeskaja pererabotka teksta. Tematičeskij sbornik. Kišinev: Štiinca, 1972, s. 117-131.

9. Kaširina M.E. O tipach raspredelenija leksičeskich edinic v tekste. StRAAT, 1974, s. 335-360.

10. Kokočašvili T.G. Raspredelenie lingvističeskich edinic v tekstach (na materiale gruzinskich i anglijskich naučno-techniceskich tekstov). Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata filologičeskich nauk. Leningrad: LGU, 1990. 16 s.

11. Kondrat'eva Ju.N., Sokolova S.V. Principy organizacii algoritmov mašinnogo perevoda. StRAPT, 1988, s. 42-50.

12. Koncepcija Gosudarstvennoj naučno-techničeskoj programmy "Perspektivnye informacionnye technologii". NTI. Serija 1, 1990, N 8, s. 2-17.

13. Lurija A.R. Jazyk i soznanie. Moskva: Izdatel'stvo Moskovskogo universiteta, 1979, 319 s.

14. Mel'nikov G.P. Sistemologija i jazykovye aspekty kibernetiki. Moskva: Sovetskoe radio, 1978, 367 s.

15. Mel'čuk I.A., Ravič R.D. Avtomatičeskij perevod. 1949-1963. Kritiko-bibliografičeskij spravočnik. Moskva: VINITI - Institut jazykoznanija AN SSSR, 1967, 517 s.

16. Metody avtomatičeskogo analiza i sinteza teksta. Piotrovskij R.G., Bilan V.N., Borkun M.N., Bobkov A.K. Minsk: Vyšejšaja škola, 1985, 222 s.

17. Muzalevskaja V.M. Vosproizvodjaščaja inženerno-lingvističeskaja model' logiko-smyslovogo analiza i perevoda (na materiale anglo-amerikanskich patentov po technologii proizvodstva pečatnych plat). Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata filologičeskich nauk. Moskva: Voennyj Krasnoznamennyj institut, 1988, 19 s.

18. Opyt mašinnnogo perevoda anglijskich i japonskich naučnych tekstov. Bektaev K.B., Beljaeva L.N., Koroleva L.N., Marčuk, Ju.N., Piotrovskij, R.G., Sadčikova, P.V., Čajkovskaja I.I., Šingareva E.A. NTI ser. 2, 1982, N 5, s. 26-31.

19. Osnovy statističeskoj optimizacii prepodavanija inostrannych jazykov. Alekseev P.M., German-Prozorova L.P., Piotrovskij R.G., Sčepetova O.P. StRAAT, 1974, Leningrad: Nauka 1974, s. 195-234.

20. Piotrovskaja K.R. Sovremennaja komp'juternaja lingvodidaktika. NTI, serija 2, 1991.

21. Piotrovskij R.G. Inženernaja lingvistika i lingvističeskie avtomaty. Vestnik AN SSSR, 1978, N 9, s. 112-119.

22. Piotrovskij R.G., Tekst, mašina, čelovek. Leningrad: Nauka LO, 1975. 327 s.

23. Piotrovskij R.G., Bektaev K.B., Piotrovskaja A.A. Matematičeskaja lingvistika. Moskva: Vyšsaja škola, 1977, 383 s.

24. Popeskul A.N. Produkcionno-setevoj podchod k modelirovaniju smysla naučno-techničeskogo teksta. Avtoreferat dissertacii na soiskanie učenoj stepeni doktora techničeskich nauk. Vinnica: Politechničeskij institut, 1991, 37 s.

25. Prekup A.V. Modelirovanie i realizacija lingvističeskogo avtomata stratifikacionnogo tipa. Avtoreferat dissertacii na soiskanie učenoj stepeni kandidata techničeskich nauk. Kišinev: Politechničeskij institut, 1989. 17 s.

26. Čižakovskij V.A. Semiotiko-kommunikativnye aspekty avtomatičeškoj pererabotki zagolovka naučno-techničeskogo teksta. Avtoreferat dissertacii na soiskanie učenoj stepeni doktora filologičeskich nauk. Leningrad: LGU, 1988. 31 s.

27. Čižakovskij V.A., Bektaev K.B. Statistika reči. 1957-1985. Bibliografičeskij ukazatel'. Kišinev: Štiinca, 1986. 111s.

28. Šingareva E.A. Semiotičeskie osnovy lingvističeskoj informatiki. Leningrad: LGPI im. A.I. Gercena, 1987. 82 s.

29. Šingareva E.A. Pragmatika v sisteme FRANT (teorija i praktika). StRAPT, 1988, s. 93-114.

30. Bailin A. Artificial Intelligence and Computer-Assisted Language Instruction: A Perspective. CALICO Journal. Computer Assisted Language Learning & Instruction Consortium. Vol.5, no 3, 1988. Pp. 25-50.

31. Beliaeva L.N., Kondratjeva Ju., Piotrowski R. & Sokolova S. Abstracts from the Leningrad MT Project. Society for Conceptual and Content Analysis by Computer. ed. by K.M. Schmidt, R.A. Boggs et al. Newsletters, N 5, Bowling Green (OH): Bowling Green State University, - Mannheim (Germany): ZUMA - Tampa, FL: The University of South Florida, 1989/90. Pp. 26-35.

32. Best K.H., Beoethy E., Altmann G. Ein methodischer Beitrag zum Piotrowski-Gesetz. Hammerl R. (ed.) Glottometrika 12. Bochum: N. Brockmeyer, 1990. S. 115-124.

33. Gardner H. The Mind's New Science. A History of the Cognitive Revolution. With a New Epilogue by the Author: Cognitive Science After 1984. N.Y.: Basic Books, Inc., 1985-87. 430 c.

34. Kuhns R.J., Little A.D. News Analysis: A Natural Language Application to Text Precessing. American Association for Artificial Intelligence. Spring Symposium Series, Text-Based Intelligence Systems. Palo Alto CA: Stanford University, 1990. Pp. 120-123.

35. Loatman R.B., Post S.D. A Natural Language Processing System for Intelligence Message Analysis. Signal. Vol.42, no 1, 1988. Pp. 41-45.

36. Piotrowski R. Linguistic Automaton and Computer Assisted Language Learning via the Microcomputer. Proceedings of CALL'89. International Conference on Computer-Assisted Language Learning at the Institute of Applied Linguistics Wilhelm Pieck University Rostock. 15-17 November 1989. Rostock: University of Rostock, 119 S.

37. Piotrowski R., Popeskul A., Chažinskaja M., Rachubo N. Automatische Wortschatzanalyse. - Bochum: N. Brockmeyer, 1985. 187 S.

38. Piotrowski R., Lesohin M., Lukjanenkov K. Introduction of Elements of Mathematics to Linguistics. Bochum: N. Brockmeyer, 1990. 260 p.

39. Piotrowski R., Grigorieva T., Piotrowska Ks., Ovsyanikov A., Gogolinskaya N. Didactic Linguistic Automaton (DLA). Proceedings of the International Conference on Microelectronics and Computer Science. ICMCS-91. Kishinev, Moldova, October 21 - 23, 1992. Vol. 3. Computer systems and image processing (A.M. Popescu, V.L. Perzhu and S.L. Rotari eds.). Kishinev Politechnical Institute, 1992. Pp. 1-8.

Glottometrika 14, 1993, 190-196

A model for the distribution of syllable types*

P. Zörnig, G. Altmann, Bochum

1. Whatever the classification of the sounds of a language may be, in syllables it is always possible to determine a nucleus and to delimit the periphery (onsets and codas) by certain criteria. The nucleus can be a vowel, a diphthong or a "syllabic" consonant, while the periphery consists essentially of consonants. To simplify matters we denote the sounds/phonemes/segments forming the nucleus by V, those forming the periphery by C. In this connection it is not relevant what a syllable "is", i.e. to which phonetic level it belongs (phonetics, phonology etc.) or how it comes into existence. The only important fact is that the syllable is a real linguistic entity participating in the dynamics of the self-regulation of language. This participation is demonstrated by the relations of the properties of syllables to other properties of language:

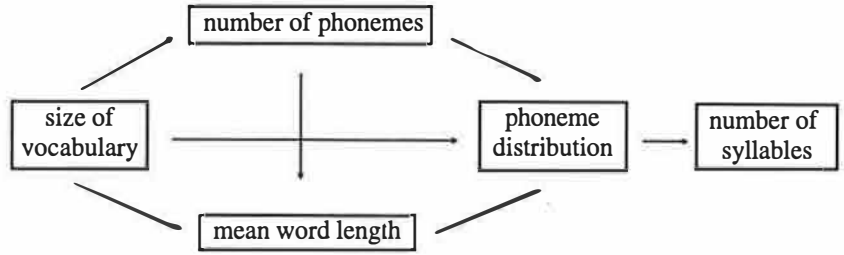
(a) The *length* of syllables depends stochastically on the length of the words in which they occur. The corresponding hypothesis is called "Menzerath's law" (cf. Altmann 1980).

(b) The *number* of syllables formed in a language depends on the number of phonemes, the number of words in the word stock, the sound/phoneme distribution and the average length of the words in the language, i.e. a large vocabulary needs many phonemes and/or a richer phoneme distribution and/or longer words. The functional equivalents can be chosen according to an optimum state for any language individually. Moreover, the number of phonemes in the inventory affects the phoneme distribution and word length: A small number of phonemes needs a richer distribution and/or longer words (or a functional equivalent such as tone or accent, which is not considered here). With a richer phoneme distribution the words can become shorter, but the number of syllables increases. The corresponding hypothesis is part of the self-regulation scheme of Köhler (cf.

* This study was written as a part of the project "Language Synergetics". The authors are grateful to the STIFTUNG VOLKSWAGENWERK for the kind support.

Köhler 1986). The net of dependencies can be schematically represented as follows:

(c) The *construction* of syllables abides by some quantitative regularities. These



may differ from one language to another, but they possibly obey some least-effort-laws (cf. e.g. Vennemann 1982; Lee 1986).

(d) The *number of canonical syllable constructs* (CV, VC, CVC, ...) is neither chaotic, nor arbitrary, nor deterministic, but obeys a stochastic distribution law, the investigation of which is subject of this paper.

2. In order to motivate our approach we make the following agreements:

Let I, J be random variables, where i and j denote the number of consonants before and after the nucleus.

The probability of i consonants before and j consonants after the nucleus is denoted by P_{ij}, the corresponding number by n_{ij}. For example n_{12} and n_{32} are the numbers of syllables of type CVCC and CCCVCC.

Moreover we make the following assumptions:

(1) At present we can investigate only the Indonesian language, since data from other languages are not known to us. Thus the theoretical result will be rather specific. Further languages may turn out to be simplified special cases or even generalizations of our model, which has a tentative character only and is meant to encourage further investigation.

(2) Every language prefers one or more special syllable types (most usually the types CV and CVC), the probability of which must be weighted in some cases.

(3) The canonical classes are formed proportionally to the classes with lower values of the random variables, i.e.

$$P'_{ij} \sim P'_{i,j-1}$$
$$P'_{ij} \sim P'_{i-1,j}$$

If we are right, then the problem to be solved is to find the correct proportionality between the individual classes. Let us consider the two-dimensional empirical distribution of Indonesian syllable types represented in Table 1. The data result from about 15000 Indonesian word forms from different texts. We suppose that further types exist, but the increase in the cells of Table 1 is assumed to be proportional, so the model should remain valid for additional data.

Table 1: Frequency distribution of Indonesian syllable types

	V	VC	VCC	VCCC
V	6	36	7	-
CV	36	391	44	2
CCV	9	61	13	-
CCCV	1	4	-	-

Table 1 can be represented in the form of Table 2.

Table 2: General form of Table 1

	0	1	2	...	j	...	k
0	n_{00}	n_{01}	n_{02} ...		n_{0j} ...		n_{0k}
1	n_{10}	n_{11}	n_{12}	...	$n_{1.j}$	...	n_{1k}
2	n_{20}	n_{21}	n_{22}	...	$n_{2.j}$	...	n_{2k}
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
i	n_{i0}	n_{i1}	n_{i2}	...	$n_{i.j}$	...	n_{ik}
.	.	.	.		.		.
.	.	.	.		.		.
r	n_{r0}	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rk}

Since the size of the syllable periphery cannot be infinite, we truncate the distribution; in our special case we set $n_{ij} = 0$ for $i > 4$ or $j > 4$.

For the proportionality between nondiagonal neighbouring classes we propose

$$P'_{ij} = \frac{b}{j^m} P'_{i,j-1} \qquad (1a)$$

$$P'_{ij} = \frac{a}{i^k} P'_{i-1,j} \qquad (1b)$$

i.e. every class is proportional to the left or upper neighbour respectively. For $i > 1$ or $j > 1$ the values decrease considerably, so we assume an inverse power-proportionality (to the value of the variable achieved). This assumption is very general ($m,k \in \mathbf{R}$). For individual languages the model can possibly be simplified.

Solving the equations (1) simultaneously leads to

$$P'_{ij} = \frac{a^i b^j}{(i!)^k (j!)^m} P'_{00} \quad i,j = 0,1,...,4 \qquad (2)$$

where P'_{00} results from the normalization:

$$P'_{00} = \left[\sum_{i=0}^{4} \sum_{j=0}^{4} \frac{a^i b^j}{(i!)^k (j!)^m} \right]^{-1} .$$

Up to the truncation, this distribution can be considered as a generalization of a model from queueing theory (cf. Cooper 1981: 127).

According to assumption (2), a modification of one or more classes may be necessary. In the case of the Indonesian language n_{11} deviates considerably from the neighbours. This class seems to be the preferred one and is weighed by β. Because of the normalization all other classes must be weighted with α following from $\beta P'_{11} + \alpha (1 - P'_{11}) = 1$. Thus we obtain the model

$$P_{ij} = \begin{cases} \beta P'_{ij} & \text{for } i = j = 1 \\ \alpha P'_{ij} & \text{for } i,j = 0,1,...,4, \text{ if } i \neq 1 \text{ or } j \neq 1 \end{cases} \qquad (3)$$

where the probabilities P'_{ij} are given by (2). For all possible indices i, j ($0 \leq i, j \leq 4$) the probabilities P_{ij} are represented in Table 3.

Table 3: Theoretical probabilities in (3)

$\alpha P'_{00}$	$\alpha b P'_{00}$	$\frac{\alpha b^2}{(2!)^m} P'_{00}$	$\frac{\alpha b^3}{(3!)^m} P'_{00}$	$\frac{\alpha b^4}{(4!)^m} P'_{00}$
$\alpha a P'_{00}$	$\beta a b P'_{00}$	$\frac{\alpha a b^2}{(2!)^m} P'_{00}$	$\frac{\alpha a b^3}{(3!)^m} P'_{00}$	$\frac{\alpha a b^4}{(4!)^m} P'_{00}$
$\frac{\alpha a^2}{(2!)^k} P'_{00}$	$\frac{\alpha a^2 b}{(2!)^k} P'_{00}$	$\frac{\alpha a^2 b^2}{(2!)^k (2!)^m} P'_{00}$	$\frac{\alpha a^2 b^3}{(2!)^k (3!)^m} P'_{00}$	$\frac{\alpha a^2 b^4}{(2!)^k (4!)^m} P'_{00}$
$\frac{\alpha a^3}{(3!)^k} P'_{00}$	$\frac{\alpha a^3 b}{(3!)^k} P'_{00}$	$\frac{\alpha a^3 b^2}{(3!)^k (2!)^m} P'_{00}$	$\frac{\alpha a^3 b^3}{(3!)^k (3!)^m} P'_{00}$	$\frac{\alpha a^3 b^4}{(3!)^k (4!)^m} P'_{00}$
$\frac{\alpha a^4}{(4!)^k} P'_{00}$	$\frac{\alpha a^4 b}{(4!)^k} P'_{00}$	$\frac{\alpha a^4 b^2}{(4!)^k (2!)^m} P'_{00}$	$\frac{\alpha a^4 b^3}{(4!)^k (3!)^m} P'_{00}$	$\frac{\alpha a^4 b^4}{(4!)^k (4!)^m} P'_{00}$

Obviously the model contains too many parameters and must be tested for many languages.

3. The easiest estimation of the parameters can be performed by means of the frequency classes as follows:

Computing the ratios of the respective P_{ij} yields

$$a^* = n_{10}/n_{00} \qquad (4a)$$

$$b^* = n_{01}/n_{00} \qquad (4b)$$

$$k^* = \ln(a^* n_{10}/n_{20})/\ln 2 \qquad (4c)$$

$$m^* = \ln(b^* n_{01}/n_{02})/\ln 2 \qquad (4d)$$

From $P_{11}/P_{10} = \beta\, b/\, \alpha$ follows

$$\beta^* = \alpha\, (^*n_{11})/(n_{10} b^*). \qquad (4e)$$

Since the P'_{ij} represent a probability distribution, we obtain

$$P'_{11} + \sum_i \sum_j P'_{ij} = 1 \quad (i \neq 1 \ \text{or} \ j \neq 1)$$

From this follows by (3):

$$\frac{P_{11}}{\beta} + \sum_i \sum_j \frac{P_{ij}}{\alpha} = 1 \quad (i \neq 1 \ \text{or} \ j \neq 1)$$

$$\Rightarrow \frac{P_{11}}{\beta} + \frac{1-P_{11}}{\alpha} = 1$$

Substituting β^* from (4e) in the above formula yields

$$\frac{P_{11}\, n_{10}\, b^*}{\alpha^*\, n_{11}} + \frac{1 - P_{11}}{\alpha^*} = 1.$$

Taking the estimation $P^*_{11} = n_{11}/N$ ($N = \Sigma\Sigma n_{ij}$) we obtain:

$$\alpha^* = 1 + \frac{n_{10}\, b^* - n_{11}}{N}.$$

For Indonesian the above parameters are as follows:

$a^* = b^* = 36/6 = 6;$
$k^* = \ln[6(36/9)]/\ln 2 = 4.585;$
$m^* = \ln[6(36/7)]/\ln 2 = 4.948;$
$\alpha^* = 1 + [36(6) - 391]/610 = 0.713;$
$\beta^* = 0.713(391)/[36(6)] = 1.291;$
$P'_{00} = 0.014.$

Thus we obtain

$$P_{ij} = \begin{cases} 0.713\, P'_{ij} & \text{for } i \neq 1 \text{ or } j \neq 1 \\ 1.291\, P'_{ij} & \text{for } i = j = 1 \end{cases}$$

with

$$P'_{ij} = \frac{(0.014)\, 6^{i+j}}{(i!)^{4.585}\, (j!)^{4.948}} \quad (0 \leq i, j \leq 4).$$

The corresponding theoretical frequencies are computed in Table 4.

Tab. 4: Fitting the distribution (3) to the Indonesian data

	0	1	2	3
0	6.1	36.5	7.1	0.2
1	36.5	396.9	42.6	1.1
2	9.1	54.8	10.7	0.3
3	0.4	2.1	0.4	0.0

The values could be further improved by optimization, but obviously the fitting is acceptable without test.

4. This model should be considered as a first approximation. It shows that the construction of syllable types follows an obviously nomological pattern according to which the number of longer types changes proportionally to that of the shorter ones. Data from other languages are necessary in order to test the adequacy of the model.

References

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

Cooper, R. B. (1981). *Introduction to queueing theory.* New York, Macmillan.

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

Lee, Sang-Oak (1986). An explanation of syllable structure change. *Korean Language Research 22, 195-213.*

Vennemann, T. (1982) (ed.). *Zur Silbenstruktur der deutschen Standardsprache. Silben, Segmente, Akzente.* Tübingen, Niemeyer, 261-305.

Glottometrika 14, 1993, 197-212

Bibliography of quantitative research into Finnish and other Finno-Ugric languages in Finland

Raimo Jussila & Anna-Liisa Kristiansson-Seppälä, Helsinki

Alm, Kaarina - Klaavu, Tuula - Haipus, Marjatta (1976). *Substantiivit 1960-luvun suomen aikakaus- ja sanomalehdissä sekä tietokirjallisuudessa* (Nouns in the Finnish periodicals, newspapers and non-fiction of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 5). 55 p.

Antila, Ulla (1982). Jälkitavujen a, ä -loppuiset vokaaliyhtymät (Vowel combinations ending in a, ä in non-initial syllables). In: *Tampereen puhekieli tutkimuskohteena* (= Folia Fennistica & Linguistica Vol. 6). Tampere, Tampereen yliopisto: 83-95.

Branch, Michael - Niemikorpi, Antero - Saukkonen, Pauli (1980). *A student's glossary of Finnish: the literary language arranged by frequency and alphabet: English - French - German - Hungarian - Russian - Swedish.* Helsinki, WSOY. 378 p.

Fiilin, Ullamaija (1980). Lexikal täthet och variation i tvåspråkiga helsingforsares finska och svenska intervjusvar (Lexical frequency and variation in Finnish and Swedish interview answers by bilingual inhabitants of Helsinki). In: Saari, M., Solstrand, H. (eds.), *Xenia Thorsiana: en vänskrift tillägnad Carl-Eric Thors på hans 60-årsdag den 8 juni 1980.* (= Meddelanden från institutionen för nordiska språk och nordisk litteratur vid Helsingfors universitet Vol. 5.) Helsinki, Helsingin yliopisto: 46-60.

Flint, Aili (1980). *Semantic Structure in the Finnish Lexicon: Verbs of Possibility and Sufficiency.* Helsinki, SKS (= Suomalaisen Kirjallisuuden Seuran toimituksia Vol. 360). 220 p.

Forsman Svensson, Pirkko (1985). Partisiippi- ja translatiivirakenteen taajuudesta ja matriisiverbeistä vanhassa kirjasuomessa ja nykysuomessa (On the

frequency of participial and translative structures and on matrix verbs in Old written Finnish and Modern Finnish). *Virittäjä (Helsinki) 89, 57-64.*

Haikola, Kaija (1956). Taivutussijojemme yleisyystilastoa lounais- ja satakuntalaismurteista (Frequency statistics of Finnish cases in south-western and Satakunta dialects). *Virittäjä (Helsinki) 60, 396-400.*

Haipus, Marjatta (1975). *Koordinaatio 1960-luvun suomen yleiskielessä (Coordination in the standard Finnish of the 1960's).* Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 4). 53 p.

Haipus, Marjatta - Pääkkönen, Helvi (1980). *Nykysuomen denominaalisista verbijohdoksista* (Denominal verb derivatives in current Finnish). Oulu, Oulun yliopisto. (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 19). 84 p.

Hakulinen, Auli - Karlsson, Fred - Vilkuna, Maria (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus* (Features of Finnish sentences in text: a quantitative investigation). Helsinki, Helsingin yliopisto (= Department of General Linguistics University of Helsinki Publications Vol. 6). 189 p.

Hankonen, Ritva (1958). Taivutussijojen yleisyystilastoa eteläpohjalaismurteesta (Frequency statistics of cases in south Ostrobothnian dialects). *Virittäjä (Helsinki) 68, 209-211.*

Henttonen, Veli-Pekka (1982). i:n loppuheitto (Apocope of i). In: *Tampereen puhekieli tutkimuskohteena* (= Folia Fennistica & Linguistica Vol. 6). Tampere, Tampereen yliopisto: 61-81.

Hienonen, Mirja (1976). Adjektiivisen predikatiivin sijasta infinitiivin yhteydessä (On the case of the predicate adjective in connection with the infinitive). In: Lehtinen, R., Lehtinen, T., Nuolijärvi, P., Paunonen, H. (eds.), *Kielitieteellisiä lehtiä* (= Suomi 120:4). Helsinki, SKS: 28-35.

Huttunen, Leena (1986). Vierassanat Jyväskylän puhekielessä (Foreign words in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 4.* Jyväskylä, Jyväskylän yliopisto: 33-57.

Huttunen, Leena (1986). sti-loppuiset adverbit ja intensiteettiadverbit jyväskyläläisessä puhekielessä (Adverbs ending sti and adverbs of intensity in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osa-*

tutkimus. Raportti 4. Jyväskylä, Jyväskylän yliopisto: 59-97.

Hyvönen, Tuula - Jämsä, Tuomo (1978). *Tempukset 1960-luvun suomen lehti- ja puhekielessä* (Tenses in the Finnish press language and informal standard speech of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 13). 28 p.

Häkkinen, Kaisa (1977). Tilastotietoja suomen kielen äännerakenteesta (Statistical information about the phonemic structure of Finnish). *Sananjalka (Turku) 19, 57-68.*

Häkkinen, Kaisa (1982). Suomen kielen sanaston suomalais-ugrilaiset juuret (The Finno-Ugric roots of the Finnish vocabulary). *Sananjalka (Turku) 24, 7-23.*

Häkkinen, Kaisa (1982). Statistische Angaben zur Lautstruktur der finnischen Sprache. *Finnisch-ugrische Mitteilungen (Hamburg) 6, 77-96.*

Häkkinen, Kaisa (1985). *Suomen kielen vanhimmasta sanastosta ja sen tutkimisesta. Suomalais-ugrilaisten kielten etymologisen tutkimuksen perusteita ja metodiikkaa* (On the oldest Finnish vocabulary and its research. The bases and methodology of the etymological research of the Finno-Ugric languages). Turku, Turun yliopisto (= Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja Vol. 17). 447 p.

Häkkinen, Kaisa (1992). Suomen perussanaston etymologiset kerrostumat (The etymological strata of the Finnish lexicon). *Virittäjä (Helsinki) 96, 47-59.*

Ikola, Osmo (1979). Hiukan sanaluokkien yleisyyssuhteista suomen murteissa ja kirjakielessä (Statistical analysis of the word classes in the Finnish dialects and written language). In: Schiefer, E.F. (ed.), *Explanationes et tractationes Fenno-Ugricae in honorem Hans Fromm.* (= Münchener Universitäts-Schriften Vol. 3). München, Wilhelm Fink Verlag: 79-86.

Ikola, Osmo (1985). Puhutun ja kirjoitetun kielen välisistä syntaktisista eroista (Syntaktische Unterschiede zwischen gesprochener und geschriebener Sprache). *Sananjalka (Turku) 27, 33-44.*

Ikola, Osmo - Palomäki, Ulla - Koitto, Anna-Kaisa (1989). *Suomen murteiden lauseoppia ja tekstikielioppia* (Syntax and Text Grammar in Finnish Dialects). Helsinki, SKS (= Suomalaisen Kirjallisuuden Seuran toimituksia Vol. 511). XIX + 568 p.

Jonninen-Niilekselä, Kaija (1982). Eräitä äänne- ja muoto-opillisia piirteitä (Some phonological and morphological features). In: *Tampereen puhekieli tutkimuskohteena* (= Folia Fennistica & Linguistica Vol. 6). Tampere, Tampereen yliopisto: 121-160.

Jussila, Raimo (1988). Agricolan sanasto ja nykysuomi (Agricola's vocabulary and modern Finnish). In: Koivusalo, E. (ed.), *Mikael Agricolan kieli.* Helsinki, SKS (= Tietolipas 112): 203-227.

Jussila, Raimo (1989). Leksikografinen luokittelu kvantitatiivisesta näkökulmasta (Lexicographical classification from the quantitative point of view). In: *XV kielitieteen päivät Oulussa 13. - 14.5.1988.* Oulu, Oulun yliopisto (= Acta Universitatis Ouluensis. Series B. Humaniora Vol. 14): 85-91.

Jussila, Raimo (1989). Suomen yleiskielen ja murteiden sanastojen suhteista (Finnish standard and dialectical lexicons compared). *Virittäjä (Helsinki) 93, 309-321.*

Jussila, Raimo (1990). Kiihtelysvaaran murteen sanasto ja yleiskieli (Vocabulary of Kiihtelysvaara dialect and standard language). In: Suni, H. (ed.), *Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10.1990.* Helsinki, VAPK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja Vol. 60): 40-60.

Jussila, Raimo - Nikunen, Erja - Rautoja, Sirkka (1992). S*uomen murteiden taajuussanasto. A Frequency Dictionary of Finnish Dialects.* Helsinki, VAPK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja Vol. 66). XXIX + 319 p.

Jussila, Raimo - Nikunen, Erja - Rautoja, Sirkka (1992). *Suomen murteiden taajuussanasto. Yksiesiintymäisten sanojen luettelo. A Frequency Dictionary of Finnish Dialects. The Listing of Hapax legomena.* Helsinki, Kotimaisten kielten tutkimuskeskus. 32 p.

Jussila, Raimo - Saukkonen, Pauli - Tuomi, Tuomo (1989). Quantitative Lexicology of Finnish. *Glottometrika 10, 155-170.*

Järvi, Martti (1981). Murrepiirteet eteläpohjalaisten opiskelijoiden ja koululaisten puhekielessä: konsonantit (Dialect features in the spoken language of Southern Ostrobothnian students and school children: consonants). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 2.* Jyväskylä, Jyväskylän yliopisto: 9-56.

Järvi, Martti (1981). Murrepiirteet eteläpohjalaisten opiskelijoiden ja koululaisten puhekielessä: vokaalit (Dialect features in the spoken language of Southern Ostrobothnian students and school children: vowels). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 2.* Jyväskylä, Jyväskylän yliopisto: 57-90.

Järvikoski, Olli (1978): Tekstin hahmottamisen dimensioista - tietokonelingvistinen sovellus (On the dimensions of text perception: an computational linguistic application). In: Alhoniemi, A. et al. (eds.), *Rakenteita. Juhlakirja Osmo Ikolan 60-vuotispäiväksi 6.2.1978* (= Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja Vol. 6). Turku, Turun yliopisto: 323-343.

Järvikoski, Olli (1985). Suomen kielen foneemien ja grafeemien frekvensseistä (On the frequencies of Finnish phonemes and graphemes). *Virittäjä (Helsinki) 89, 33-47.*

Kajava, Jorma - Kekäläinen, Jorma (1981). First- and second-order entropies and redundancies of Finnish language. *Suomalais-ugrilaisen seuran aikakauskirja (Helsinki) 77, 275-279.*

Kajava, Kalevi (1962). Alkeis- ja oppikirjojen sanastontutkimuksesta (De l'etude du vocabulaire des manuels elementaires). *Virittäjä (Helsinki) 66, 20-33.*

Karlsson, Fred (1983). *Suomen kielen äänne- ja muotorakenne* (Phonological and morphological structure of Finnish). Helsinki, WSOY: 410 p.

Karlsson, Göran (1966). Eräitä tilastollisia tietoja subjektin ja predikaatin numeruskongruenssista suomen murteissa (Einige statistische Angaben über die Numeruskongruenz des Subjekts und des Prädikats in den finnischen Mundarten). *Sananjalka (Turku) 8, 17-23.*

Karvonen, Juhani - Takala, Annika - Röman, Kyllikki & Ylinentalo, Oiva (1970). *Opettajan sanastokirja* (Teacher's vocabulary book). Jyväskylä, Gummerus.

Kiuru, Silva (1988). Agricolan teonnimijohdosten erikoispiirteitä (Action noun derivatives in -mus, -mys in Agricola's language). In: Koivusalo, E. (ed.), *Mikael Agricolan kieli.* Helsinki, SKS (= Tietolipas 112): 133-179.

Kiuru, Silva (1988). Agricola tulepi, ios henen tule. Ind. preesensin yks. 3. persoonan muodot Mikael Agricolan kielessä (The 3rd person singular present indicative in Agricola's language). In: Kalliokoski, J., Leino, P., Pyhtilä, P.

(eds.), *Kieli 3*. Helsinki, Helsingin yliopiston suomen kielen laitos: 7-76.

Kiviniemi, Eero (1982). *Rakkaan lapsen monet nimet. Suomalaisten etunimet ja nimenvalinta* (The many names of a dear child: the given names of the Finns and their selection). Espoo, Weilin & Göös. 376 p.

Kiviniemi, Eero (1990). Kajahtiko Karjalasta?: karjalaisten etunimimieltymykset tilastojen valossa (Statistical study on Karelian given names). In: Suni, H. (ed.), *Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10. 1990*. Helsinki, VAPK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja Vol. 60): 78-93.

Koistinen, Anneli - Jämsä, Tuomo - Haipus, Marjatta (1976). *Deverbaaliset substantiivit 1960-luvun suomenkielisessä kauno- ja tietokirjallisuudessa* (Deverbal nouns in the Finnish fiction and non-fiction of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 8). 31 p.

Koivusalo, Esko (1980). *Index Agricolaensis I - II*. Helsinki, KKTK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja 11). 879 p.

Koivusalo, Esko - Suni, Helena (1988). Autuahus taiuahisa. Jälkitavujen vokaalien välinen h Agricolan teosten kielessä (h between vowels after syllables with main stress in Agricola's language). In: Koivusalo, E. (ed.), *Mikael Agricolan kieli*. Helsinki, SKS (= Tietolipas 112): 111-132.

Korhonen, Mikko (1969). Die Entwicklung der morphologischen Methode Im Lappischen. *Finnisch-ugrische Forschungen (Helsinki) 37, 203-362.*

Koskinen, Juhani (1965). Tilastollisia havaintoja Mikael Agricolan aktiivin 2. partisiipista (Statistical observations on Mikael Agricola's 2nd active participle). *Virittäjä (Helsinki) 69, 172-175.*

Kurikka, Ulla (1979). *Kielto 1960-luvun suomen yleiskielessä* (Negation in the standard Finnish of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 17). 61 p.

Kurkkio, Markku (1978). *Referaatti 1960-luvun suomen yleiskielessä* (Reported speech in the standard Finnish of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 16). 66 p.

Kuusi, Matti (1972). Savolaissuvuista (On Savo clans). In: Sihvo, H. (ed.), *Nimikirja*. Helsinki, Kalevalaseura (= Kalevalaseuran vuosikirja Vol. 52): 99-115.

Laalo, Klaus (1982). Nykysuomen nominivartalotyyppien ja niiden ikäkerrostumien frekvensseistä: kaksitavuisten nominien eri vartalotyyppien sanamäärät ja sanojen esiintymistaajuudet (On the frequencies of nominal stem types and their age strata in Modern Standard Finnish). *Virittäjä (Helsinki) 86, 22-42.*

Lahtinen, Leena (1981). a:n ja ä:n loppuheitto Jyväskylän puhekielessä (Apocope of a and ä in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 3*. Jyväskylä, Jyväskylän yliopisto: 35-50.

Laurinen, Inkeri (1955). *Lausetajun kehityksestä suomenkielisen kansakoulun kirjoituksenopetuksen tulosten valossa* (The development of sentence sense in the light of the results attained in teachning of writing in Finnish primary schools). Turku. 276 p.

Lehtiranta, Juhani (1982). Eine Beobachtung über die Gründe der raschen Veränderung des Grundwortschatzes im Lappischen. *Finnisch-ugrische Forschungen (Helsinki) 44, 114-118.*

Leino, Pentti (1970). *Strukturaalinen alkusointu suomessa: folklorepohjainen tilastoanalyysi* (Structural alliteration in Finnish.) Helsinki, SKS (= Suomalaisen Kirjallisuuden Seuran toimituksia Vol. 298). 322 p.

Leino, Pentti (1971). Norjanlappalaisen sananparsiston alkusointuisuus (Alliteration in the sayings of the Norwegian Sami). In: Sihvo, H. (ed.), *Vanhaa ja uutta lappia* (= Kalevalaseuran vuosikirja 51). Helsinki, WSOY: 178-188.

Leino, Pentti (1972). Etunimien suosionvaihteluja (Variations in the popularity of Finnish given names). In: Sihvo, H. (ed.), *Nimikirja*. Helsinki, Kalevalaseura (= Kalevalaseuran vuosikirja Vol. 52): 75-89.

Leiwo, Matti (1968). Testausmenettelyistä merkitysopillisissa tutkimuksissa (Testing methods in semantic studies). *Virittäjä (Helsinki) 72, 151-158.*

Lepistö, Eino (1938). Vampulan murteen äänteiden yleisyystilastoa (Frequenzstatistik der Laute in der Mundart von Vampula). *Virittäjä (Helsinki) 42, 45-53.*

Leppäjärvi, Eila - Jämsä, Tuomo (1976). *Adjektiivit 1960-luvun suomen lehti-, radio- ja yleispuhekielessä* (Adjectives in the Finnish press and radio language and informal standard speech of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 7). 32 p.

Leskinen, Heikki (1971). Tilastollisia havaintoja kaakkoismurteiden loppu-n:stä (Statistische Beobachtungen über den Schwund des auslautenden -n in den Südostdialekten des Finnischen). *Virittäjä (Helsinki) 75, 343-368.*

Leskinen, Heikki (1973). Kaakkoissuomalaisen loppuheiton yleisyydestä ja alkuperästä. (Über Allgemeinheit und Ursprung der südostfinnischen Apokope). *Suomalais-ugrilaisen seuran aikakauskirja (Helsinki) 72, 210-221.*

Leskinen, Heikki (1974). Karjalaisen siirtoväen murteen sulautumisesta ja sen tutkimisesta (Über die Verschmelzung des Dialekts der karelischen Umsiedler und dessen Untersuchung). *Virittäjä (Helsinki) 78, 361-378.*

Leskinen, Heikki (1989). Tietoja sananalkuisten grafeemien ja grafeemikombinaatioiden yleisyydestä (Daten zur Allgemeinheit der Grapheme und Graphemkombinationen im Wortanlaut des Finnischen). *Virittäjä (Helsinki) 93, 401-419.*

Leskinen, Heikki (1991). Vieläkö nuoret nurisevat? - Huomioita onomatopoeettisten sanojen tuntemuksesta ja tulkinnasta (Beobachtungen zur Kenntnis und Deutung onomatopoetischer Wörter). *Virittäjä (Helsinki) 95, 355-371.*

Leskinen, Heikki - Savijärvi, Ilkka - Särkkä, Tauno (1974). *Koululaisten kirjallisen ilmaisutaidon kehityksestä* (On the development of the skill of literary expression of school children). Jyväskylä, Jyväskylän yliopisto (= Jyväskylän yliopiston suomen kielen laitos - Julkaisuja Vol. 8). 178 p.

Matihaldi, Hilkka-Liisa (1980). *Nykysuomen modukset II. Kvantitatiivinen analyysi* (Mood in present-day Finnish). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 20). 118 p.

Mielikäinen, Aila (1978). Aktiivin II partisiippi Etelä-Savon murteissa (The past active participle in South Savo dialects). *Virittäjä (Helsinki) 82, 101-121.*

Mielikäinen, Aila (1980). Pikapuhemuodot Jyväskylän puhekielessä (Allegro speech forms in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 1.* Jyväskylä, Jyväskylän yliopisto: 123-148.

Mielikäinen, Aila (1981). Murre, kielenkäyttäjät ja asenteet (Dialect, the language user and attitudes). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 2.* Jyväskylä, Jyväskylän yliopisto: 91-126.

Mielikäinen, Aila (1981). Nominin- ja verbintaivutuksen ongelmia nykypuhekielessä (Problems of declension and conjugation in the modern spoken language). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 3.* Jyväskylä, Jyväskylän yliopisto: 67-100.

Mielikäinen, Aila (1986). Relatiivipronominit nykypuhekielessä (Relative pronouns in the modern spoken language). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 4.* Jyväskylä, Jyväskylän yliopisto: 99-126.

Mujunen, Raili - Niemikorpi, Antero - Pelttari, Maire (1979). *Pronominit 1960-luvun suomen yleiskielessä* (Pronouns in the standard Finnish of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 18). 36 p.

Nahkola, Kari - Saanilahti, Marja (1990). Lekseemin esiintymistaajuuden vaikutus kielenmuutoksen leksikaaliseen diffuusioon (On the effect of lexeme frequency on the lexical diffusion of a language change). *Virittäjä (Helsinki) 94, 196-217.*

Nahkola, Kari - Saanilahti, Marja (1991). Koululaisslangin semanttisia ja sosiolingvistisiä piirteitä (Semantic and sociolinguistic features of school slang). *Virittäjä (Helsinki) 95, 123-140.*

Niemikorpi, Antero (1983). Taajuussanasto ja ammattikielet (Frequency vocabulary and technical languages). In: *Erikoiskielet ja käännösteoria: VAKKI-seminaari 3.* Vaasa, Vaasan korkeakoulu: 21-36.

Niemikorpi, Antero (1990). *Suomen kielen sanaston frekvenssianalyysi*a (Frequency analysis of Finnish vocabulary). Vaasa, Vaasan korkeakoulu (= Vaasan korkeakoulun julkaisuja. Tutkimuksia Vol. 150). 155 p.

Niemikorpi, Antero (1991). *Suomen kielen sanaston dynamiikkaa* (Dynamics of the Finnish vocabulary). Vaasa, Vaasan yliopisto (= Acta Wasaensia 26. Universitas Wasaensis). 420 p.

Nikkilä, Osmo (1985). *Apokope und altes Schriftfinnisch. Zur Geschichte der i-Apokope des Finnischen.* Groningen. Rijksuniversiteit de Groningen. 508 p.

Nikkilä, Osmo (1988). Agricolan kieli ja teokset loppuheiton valossa (Agricola's language and work in the light of apocope). In: Koivusalo, E. (ed.), *Mikael Agricolan kieli.* Helsinki, SKS (= Tietolipas 112): 94-110.

Nissi, Ulla (1981). III infinitiivin illatiivi ja inessiivi jyväskyläläisten ja eteläpohjalaisten puhekielessä (The third infinitive illative and inessive in the vernacular of people from Jyväskylä and Southern Ostrobothnia). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 3.* Jyväskylä, Jyväskylän yliopisto: 51-65.

Nuolijärvi, Pirkko (1986). *Kolmannen sukupolven kieli. Helsinkiin muuttaneiden suurten ikäluokkien eteläpohjalaisten ja pohjoissavolaisten kielellinen sopeutuminen* (The Language of the Third Generation. The Linguistic Adaptation of Representatives of the Large Age Groups who have Moved to Helsinki from Southern Ostrobotnia and Nothern Savo). Helsinki, SKS (= Suomalaisen Kirjallisuuden Seuran toimituksia Vol. 436). 354 p.

Nuolijärvi, Pirkko (1986). "Ota minut sinun uniin" (Juice Leskinen). Nykysuomalaisen omistusmuotojärjestelmästä (On the system of possessive in modern Finnish). In: Kalliokoski, J., Leino, P., (eds.), *Kieli 1.* Helsinki, Helsingin yliopiston suomen kielen laitos: 157-182.

Olli, Raili (1986). Yleiskielestä poikkeava sanasto jyväskyläläisten ja eteläpohjalaisten puhekielessä (Non-standard vocabulary in the vernacular of people from Jyväskylä and Southern Ostrobothnia). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 4.* Jyväskylä, Jyväskylän yliopisto: 1-32.

Pajunen, Anneli - Palomäki, Ulla (1984). *Tilastotietoja suomen kielen rakenteesta 1. Frequence Analysis of Spoken and Written Discourse in Finnish 1.* Helsinki, KKTK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja 30). IV + 100 p.

Pajunen, Anneli - Palomäki, Ulla (1985). *Tilastotietoja suomen kielen rakenteesta 2. Frequence Analysis of Spoken and Written Discourse in Finnish 2.* Helsinki, KKTK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja 31). 78 p.

Palander, Marjatta (1981). Liperin murteen erikoisgeminaatio kolmen ikäpolven kielessä (Gemination in the Liperi dialect in the language of three generations). In: *Huomioita suomen murteiden geminaatioilmiöistä* (= Folia Fennistica & Linguistica Vol. 5). Tampere, Tampereen yliopisto: 83-131.

Palander, Marjatta (1987). aa:n ja ää:n diftongiutuneisuus sekä e:n labialisaatio nykysavossa (The diphthongization of a and ä and the labialization of e in modern Savo dialect). In: *Kirjoituksia kansanmurteista ja kirjakielestä* (= Folia Fennistica & Linguistica Vol. 14). Tampere, Tampereen yliopisto: 21-60.

Palander, Marjatta (1991). Puhe- ja kirjakielen sanajärjestyseroista (Word order differences in spoken and written language). *Virittäjä (Helsinki) 95, 235-254.*

Palola, Helena (1975). *Adverbit, postpositiot, prepositiot ja interjektiot 1960-luvun suomen yleiskielessä* (Adverbs, postpositions and interjections in the standard Finnish of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 2). 60 p.

Pesonen, Jaakko (1971). *Sananmuodot ja niiden kirjainrakenne suomenkielisessä sanomalehtitekstissä* (Word-inflexions and their letter and syllable structure in finnish newspaper text). Jyväskylä, University of Jyväskylä (= Department of Special Education. Research report Vol. 6/1871). 83 p.

Pitkänen, Ritva Liisa (1972). Saarta tai karia tarkoittavia maastoappellatiiveja saaristonnimissä (Natural names meaning island or rock in Finnish island names). In: Sihvo, H. (ed.), *Nimikirja.* Helsinki, Kalevalaseura (= Kalevalaseuran vuosikirja Vol. 52): 333-359.

Pääkkönen, Irmeli (1982). 5. kouluvuottaan aloittavien sanavarat (The vocabularies of school children starting their fifth school year). In: Larmola, M. (ed.), *Kouluikäisten kieli* (= Tietolipas 88). Helsinki, SKS: 110-129.

Pääkkönen, Matti (1973). Tietoja suomen yleiskielen grafeemeista (Statistische Angaben über die Grapheme der finnischen Hochsprache). *Suomalais-ugrilaisen seuran aikakauskirja (Helsinki) 72, 318-322.*

Pääkkönen, Matti (1990). *Grafeemit ja konteksti: tilastotietoja suomen yleiskielen kirjaimistosta* (Graphemes and context: statistical data on the graphology of standard Finnish). Helsinki, SKS (= Suomi Vol. 150). 129 p.

Raivio, Erkki (1974). Suomen konsonanttiyhtymien tilastollista tarkastelua

(Statistical analysis of Finnish consonants). *Virittäjä (Helsinki) 78, 186-189.*

Rantala, Tuula (1959). Eräiden raamatunsuomennosten ilmaisutiiviydestä (On the density of expressions in some Bible translations). *Virittäjä (Helsinki) 63, 394-400.*

Rautkorpi, Esko (1980). Jyväskylän murteen väistyviä piirteitä (Receding features of the Jyväskylä dialect). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 1.* Jyväskylä, Jyväskylän yliopisto: 77-88.

Roimu, Marjatta (1963). Suomen virkkeen ja sanaston rakenteesta (De la structure lexicale de la proposition du finnois). *Virittäjä (Helsinki) 67, 69-73.*

Rossi, Toini - Haipus, Marjatta (1978). *Numeraalit 1960-luvun suomen yleiskielessä* (Numerals in the standard Finnish of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 14). 32 p.

Ruoppila, Veikko (1936). Äänteiden yleisyystilastoa Lemin murteesta (Frequenzstatistik der Laute für den Dialekt von Lemi). *Virittäjä (Helsinki) 40, 127-131.*

Räsänen, Seppo (1974). "Tuuleen viritetty korva". Huomioita 1950-luvun suomalaisten runoilijoiden paikallissijojen käytöstä ("Auf den Wind gestimmtes Ohr"). *Sananjalka (Turku) 16, 16-23.*

Räsänen, Seppo (1975). *Sanastollis-kvantitatiivinen tutkimus Aleksis Kiven pääteosten tyylistä* (Lexicographical-quantitative research on the style of Aleksis Kivi's principal works). Tampere, Tampereen yliopisto (= Tampereen yliopiston suomen kielen laitos, monistesarja Vol. 4). 34 p.

Räsänen, Seppo (1979). Huomioita suomen sijojen frekvensseistä (Beobachtungen über die Häufigkeit des finnischen Kasussystem). *Sananjalka (Turku) 21, 17-34.*

Saarela, Leena (1991). *Neljäsluokkalaisten tyttöjen ja poikien sanaston eroista* (On differences in vocabulary used by boys and girls in fourth class). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 33). 47 p.

Salmelin, Annikki (1959). Kirjasuomen saneen loppuäänteiden yleisyystilastoa (Frequency statistic of the final sounds of literary Finnish words). *Virittäjä (Helsinki) 63, 400-403.*

Saukkonen, Pauli (1966). Kokeellisia havaintoja puhekielen ja kirjakielen tyylieroista (Experimentelle Beobachtungen über Stilunterschiede zwischen Umgangsprache und Schriftsprache). *Virittäjä (Helsinki) 70, 38-53.*

Saukkonen, Pauli (1967). *Svaa-analyysi savolais- ja kaakkoismurteiden rajalta* (Statistical-sociological analysis of the schwa glide phenomenon in the border area between the Savo and southeastern dialects of Finnish). Helsinki, SKS (= Suomi Vol. 111:4). 74 p.

Saukkonen, Pauli (1972). Puheen hahmotuksesta (On identifying speech). In: Vierikko, E. (ed.), *Puhekieli ja ilmaisu.* Helsinki, WSOY: 33-38.

Saukkonen, Pauli (1973). Suomen kielen yhdyssanojen rakenne. (Die Struktur der Komposita im Finnischen.) In: *Suomalais-ugrilaisen Seuran toimituksia Vol. 150.* Helsinki, SUS: 332-339.

Saukkonen, Pauli (1977). *Nykysuomen saneiston yleisyystilastoa saneenloppuisessa aakkosjärjestyksessä* (Statistics of the word occurrences of present-day Finnish in a reverse alphabetical order). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja Vol. 9). 25 p.

Saukkonen, Pauli (1980). Statistical linguistic research in Finland. (Statistiko-lingvističeskie issledovanija v Finljandii.) In: *Lingvostatistilisi uurimusi soome-ugri keelte alalt. Töid keelestatistika alalt 5* (= Tartu riikliku ülikooli toimetised 518). Tartu, Tartu ülikool: 5-14.

Saukkonen, Pauli - Haipus, Marjatta - Niemikorpi, Antero - Sulkala, Helena (1979). *Suomen kielen taajuussanasto. A frequency dictionary of Finnish.* Porvoo, WSOY. 536 p.

Savijärvi, Ilkka (1977). *Itämerensuomalaisten kielten kieltoverbi. I. Suomi* (The negative verb in the Baltic-Finnic languages. 1. Finnish). Helsinki, SKS (= Suomalaisen Kirjallisuuden Seuran toimituksia Vol. 333). 288 p.

Savijärvi, Ilkka (1977). *Redundanssi ja kieltoverbin ellipsi suomen kielen negaatiojärjestelmässä* (Redundancy and the ellipsis of the negative verb in the Finnish negation system). Jyväskylä, Jyväskylän yliopisto (= Jyväskylän yliopiston suomen kielen ja viestinnän laitoksen julkaisuja Vol. 14). 57 p.

Savijärvi, Ilkka (1988). Agricolan kieltolause (Negative sentence in Agricola's language). In: Koivusalo, E. (ed.), *Mikael Agricolan kieli.* Helsinki, SKS (= Tietolipas 112): 69-93.

Savijärvi, Ilkka (1990) Passiivin ja monikon 3. persoonan suhteesta vepsän kielessä (On passive in Vepsian). In: Suni, H. (ed.), *Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10.1990.* Helsinki, VAPK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja Vol. 60): 40-60.

Setälä, Vilho (1951). Kirjainten ja äänteiden useudet Suomessa (On the frequency of letters and sounds in Finnish). *Virittäjä (Helsinki) 55, 212-214.*

Setälä, Vilho (1967). Tilastollisia tietoja Uuden testamentin suomennoksen sanastosta (Statistical knowledge of the lexicology of the translations of the New Testament). *Virittäjä (Helsinki) 71, 368-372.*

Setälä, Vilho (1972). *Suomen kielen dynamiikkaa. 1. Kirjain- ja äännetilastoa* (Dynamics of Finnish language. 1. Statistics of graphemes and phonemes). Helsinki, SKS (= Suomi 116:3). 58 p.

Setälä, Vilho (1974). Kalevalan ja Uuden testamentin sanastosta (On the vocabulary of the Kalevala and the New Testament). In: Virtaranta, P. et al. (eds.), *Sampo ei sanoja puutu: Matti Kuusen juhlakirja* (= Kalevalaseuran vuosikirja 54). Helsinki, Kalevalaseura: 370-379.

Silvennoinen, Leena (1980). Laaja-alaiset itämurteisuudet Jyväskylän puhekielessä (Eastern features in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 1.* Jyväskylä, Jyväskylän yliopisto: 89-122.

Silvennoinen, Leena (1981). a, ä -loppuiset vokaaliyhtymät Jyväskylän puhekielessä (Vowel combinations ending in a, ä in the Jyväskylä vernacular). In: *Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Raportti 3.* Jyväskylä, Jyväskylän yliopisto: 1-34.

Suihkonen, Pirkko (1990). *Korpustutkimus kielitypologiassa sovellettuna udmurttiin* (Computer corpus analysis in language typology applied to Udmurt). Helsinki, SUS (= Suomalais-ugrilaisen Seuran toimituksia Vol. 207). 343 p.

Suojanen, Matti - Salomaa, Leena - Vuorinen, Riitta (1981). Kolmen murrepiirteen sosiaalinen ja tilanteinen erottelevuus Turun puhekielessä nykysuomalaisen puhekielen murroksen tutkimuksen osaraportti (The social and situational separation of three dialect features in the Turku spoken language an intermediate report on the research of the transition in the modern Finnish spoken language). In: Suojanen, M.K. (ed.), *Kirjoituksia puhekielestä* (= Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja 14). Turku, Turun

yliopisto: 109-143.

Särkilahti, Sirkka-Liisa (1967). Voidaanko kirjailijoiden kieltä tutkia kvantitatiivisin metodein?: Eeva Joenpellon, Vilho Niitemaan, Marja-Liisa Vartion ja F. E. Sillanpään tekstinäytteiden tilastollinen analyysi (Can an author's language be studied by quantitative methods?). In: *Juhlakirja Kauko Kyyrön täyttäessä 60 vuotta 24.11.1967* (= Acta Universitatis Tamperensis A Vol. 18). Tampere, Tampereen yliopisto: 253-266.

Särkilahti, Sirkka-Liisa (1969). Sanafrekvenssit kielen havainnollistajina (Observations on word frequences). In: Alhoniemi, A. et al. (eds.), *Juhlakirja Paavo Siron täyttäessä 60 vuotta 2.8.1969* (= Tampere Acta Universitatis Tamperensis ser. A vol. 26). Tampere, Tampereen yliopisto: 200-204.

Särkilahti, Sirkka-Liisa (1977). *Tyylintutkimuksen kvantitatiiviset metodit* (Quantitative methods in stylistic research). Turku, AFinLA (= Suomen sovelletun kielitieteen yhdistyksen (AFinLA) julkaisuja 19). 143 p.

Särkkä, Tauno (1987). Sanaston rikkaudesta ja sen mittaamisesta (Über den Reichtum des Wortschatzes und seine Messung). *Virittäjä (Helsinki) 91, 129-137.*

Tuomi, Tuomo (1989). *Suomen murteiden sanakirja. Johdanto* (Dictionary of Finnish dialects. Introduction). Helsinki, VAPK (= Kotimaisten kielten tutkimuskeskuksen julkaisuja 36). 100 p.

Vehmaskoski, Maila (1976). Sanaston frekvensseistä ja laadusta eräiden vuosina 1935 ja 1965 ilmestyneiden romaaneiden repliikeissä (On the frequency and quality of the vocabulary of the lines of some Finnish novels published in 1935 and 1965). In: Lehtinen, R., Lehtinen, T., Nuolijärvi, P., Paunonen, H. (eds.), *Kielitieteellisiä lehtiä* (= Suomi 120:4). Helsinki, SKS: 132-149.

Vierikko, Esko (1972). Nuorten poliitikkojen kielestä (On the language of the young politicians). In: Vierikko, E. (ed.), *Puhekieli ja ilmaisu.* Helsinki, WSOY: 39-48.

Virkkunen, Hilkka (1977). *Essiivi, translatiivi, abessiivi, komitatiivi ja instruktiivi 1960-luvun suomen yleiskielessä* (Essive, translative, abessive, comitative and instructive in the standard Finnish language of the 1960's). Oulu, Oulun yliopisto (= Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 1). 63 p.

Virtaranta, Pertti (1967). *F. E. Sillanpään puhetta kaksi keskustelua Sillanpään kanssa ja havaintoja hänen puheestaan* (F.E. Sillanpää talking two conversations with Sillanpää and observations on his speech). Keuruu, Otava. 104 p.

Virtaranta, Pertti (1969). Sanojen ja omistusliitteisten muotojen esiintymistiheydestä hämäläismurteissa (Zur Häufigkeit von Wörtern und possessiven Formen in den Dialekten von Häme.) In: Alhoniemi, A. et al. (eds.), *Juhlakirja Paavo Siron täyttäessä 60 vuotta 2.8.1969* (= Tampere Acta Universitatis Tamperensis ser. A vol. 26). Tampere, Tampereen yliopisto: 238-256.

Vuorenjuuri, Helena (1958). Tilastoa kansankielen sananvalinnasta ja virkkeiden muodostuksesta (Statistics on the choice of words in the Finnish dialects and the formation of sentences). *Äidinkielen opettajain liiton vuosikirja (Helsinki) 6, 95-103.*

Åstedt, Kaarina (1957). Suomen kielen taivutussijojen yleisyystilastoa (Frequency statistics of Finnish cases). *Virittäjä (Helsinki) 61, 424-427.*

***Abbreviations (publishers):**

AFinLA = Suomen sovelletun kielitieteen yhdistys, Helsinki
KKTK = Kotimaisten kielten tutkimuskeskus, Helsinki
SKS = Suomalaisen Kirjallisuuden Seura, Helsinki
SUS = Suomalais-ugrilainen Seura, Helsinki
VAPK = Valtion painatuskeskus, Helsinki
WSOY = Werner Söderström Osakeyhtiö, Helsinki

Glottometrika 14, 1993, 213

Quantitative Linguistics, by Marie Těšitelová

Praha, Academia 1992. 253 pp.

Reviewed by Luděk Hřebíček, Prague

The author belongs to the group of outstanding personalities of contemporary Czech linguistics. Her entire scientific activity has been devoted to statistical linguistics and to the description of the Czech language. The work under review can be treated as a summary of her life-time work and also as an expression of her philosophy of theis field of study. No less than 44 of her own works or of those originated with her collaboration are quoted here, including such monumental publications as the frequency dictionary of Czech published in 1961 (together with J. Jelínek and J.V. Bečka) or the Reverse Dictionary of Contemporary Czech from 1986 (in collaboration with J. Petr and J. Králík). Also, the monographs *Problems of Lexical Statistics* (1974), and *The Use of Statistical Methods in Grammar* (in Czech, 1980), and many other works are worth mentioning.

The philosophy of quantitative linguistics starts here with the ideas of classical linguistics. Quantitative linguistics is understood as an instrument for obtaining a larger repertory of characteristics of language phenomena. Frequency dictionaries of many languages are extensively commented on and characterized. Zipf 's law and the Zipf-Mandelbrot law are the only linguistic laws explained in detail. The explanation is very clear and instructive.

The book can be characterized in short as an instruction for asking the questions about WHAT and HOW. Consequently, it summarizes a certain stage in the development of linguistics which naturally will never end, as these questions will still remain relevant. However, each science, including the so-called humanities, sooner or later must reach the stage when the causes of phenomena are sought and the question WHY is asked. This stage evidently has already been reached by quantitative linguistics, which is already constructing testable theories. This is testified to by the works published in the series Quantitative Linguistics as well as by many other publications. However, in this respect the political environment must be taken into account. The previous existence of the Iron Curtain intervened at least in the section on Bibliography in the work under review, as the Curtain was impervious also to scientific publications. The greater is the respect commanded by this work.

Glottometrika 14, 1993, 214-231

CURRENT BIBLIOGRAPHY[1]

Compiled by Christiane Hoffmann, Trier

Sigla

ACS	Area and culture studies
GLOTT	Glottometrika
JEXPG	Journal of experimental psychology / General
LASCI	Language sciences
LCS	Linguistics
LIUR	Linguistica uralica
MLJAP	Mathematical linguistics. Japan
MONLAD	Mondo ladino
MUSI	Musikometrika
RLR	Revue de linguistique romane
RSSI	Recherches sémiotiques = Semiotic inquiry
STLSCI	Studies in linguistic sciences
TL	Theoretical linguistics
UAJB	Ural-altaische Jahrbücher
VCS	Vyčistel'nye sistemy (Analiz tekstov i signalov)

BABITCH 1987 = Babitch, Rose Mary (Ed.): Papers from the eleventh annual meeting of the Atlantic Provinces Linguistic Association. Centre Universitaire de Shippagan, Shippagan, New Brunswick, Nov. 13-14, 1987. 1987

GROTJAHN 1987 = Grotjahn, Rüdiger (Ed.); Klein-Braley, Christine (Ed.); Stevenson, Douglas K. (Ed.): Taking their measure. The validity and validation of language tests. Bochum: Brockmeyer, 1987 (Quantitative linguistics ; 34) ISBN: 3-88339-642-7

ROTHE 1991 = Rothe, Ursula (Ed.): Diversification processes in language. Grammar. Hagen: Rottmann Medienverl., 1991 ISBN: 3-926862-21-1

[1]Extract from the *Bibliography of Quantitative Linguistics* which is being compiled at the University of Trier. This project is supported by the Deutsche Forschungsgemeinschaft (DFG).

THOMAS 1987 = Thomas, Alan R. (Ed.): Methods in dialectology. Proceedings of the Sixth International Conference held at the University College of North Wales, 3.-7. Aug. 1987. Clevedon [u.a.]: Multilingual Matters, 1988. (Multilingual matters ; 48)

GENERAL

Altmann, Gabriel: The levels of linguistic investigation. IN: TL 14(1987), S.227-240

Anreiter, Peter: Transformierte sprachtypologische Profilvektoren. IN: GLOTT 10(1989), S.[32]-45

Arapov, Michail Viktorovič: Kvantitativnaja lingvistika. [Quantitative Linguistik]. Moskva: Nauka, 1988

Fickermann, I. (Ed.): Glottometrika 8. Bochum: Brockmeyer, 1987 (Quantitative linguistics ; 32) ISBN: 3-88339-559-5

Grotjahn, Rüdiger: Hatch, Evelyn; Farhady, Hossein: Research design and statistics for applied linguistics. Rowley, Mass., 1982. IN: GLOTT 9(1988), S.219-233 [Review]

Hammerl, Rolf: Arapov, M. V.: Kvantitativnaja lingvistika. Moskva: Nauka, 1988. IN: GLOTT 12(1990), S.189-199 [Review]

Hammerl, Rolf (Ed.): Glottometrika 12. Bochum: Brockmeyer, 1990 (Quantitative linguistics ; 45) ISBN: 3-88339-843-8

Hřebíček, Luděk (Ed.): Glottometrika 11. Bochum: Brockmeyer, 1990 (Quantitative linguistics ; 42) ISBN: 3-88339-777-6, 0932-7991

Kristophson, J.: Gliederung einer Sprachfamilie (hier der Romania) mit Hilfe eines numerischen Kalküls. IN: GLOTT 11(1990), S.[68]-94

Minagawa, Naohiro; Kashů, Kan: Haiku o kōsei suru go no sōgo kanrendo to haiku ni taisuru kyōkando to no kankei. [A relationship between the component word relatedness in a Haiku and the degree of liking for a Haiku]. IN: MLJAP 17(1990)6, S. 265-272

Mizutani, Shizuo (Ed.): Japanese quantitative linguistics. Bochum: Brockmeyer, 1989 (Quantitative linguistics ; 39) ISBN: 3-88339-723-7

Piotrovskij, Rajmund G.; Lesochin, Michail M.; Luk'janenkov, K. F.: Introduction of elements of mathematics to linguistics. Bochum: Brockmeyer, 1990 (Quantitative linguistics ; 44) ISBN: 3-88339-833-0

Schulz, Klaus-Peter (Ed.): Glottometrika 9. Bochum: Brockmeyer, 1988 (Quantitative linguistics ; 35) ISBN: 3-88339-648-6

Thümmel, Wolf: Reihenfolgebeziehungen in der syntaktischen Sprachtypologie. IN: GLOTT 9(1988), S.59-104

LANGUAGE ACQUISITION

Bachmann, Lyle F.; Palmer, Adrian S.: The construct validation of some components of communicative proficiency. IN: GROTJAHN 1987, S.91-110

Carroll, John B.: Psychometric theory and language testing. IN: GROTJAHN 1987, S.1-39

Gaies, Stephen J.: Validation of the noise test. IN: GROTJAHN 1987, S.41-74

Grotjahn, Rüdiger: How to construct and evaluate a C-Test. A discussion of some problems and some statistical analyses. IN: GROTJAHN 1987, S.219-253

Grotjahn, Rüdiger; Krause, Jürgen; Unwerth, Heinz-Jürgen von: The Bochum diagnostic test for English. IN: GROTJAHN 1987, S.133-156

Klein-Braley, Christine: Fossil at large. Translations as a language testing procedure. IN: GROTJAHN 1987, S.111-132

Liski, Erkki P.; Puntanen, Simo: Errors and the number of utterances in group conversation tests in spoken English. IN: GROTJAHN 1987, S.157-183

Raatz, Ulrich: Some remarks on the validity of diagnostic tests. IN: GROTJAHN 1987, S.75-89

Rea, Pauline M.: Testing doctor's written communicative competence. An experimental technique in English for specialist purposes. IN: GROTJAHN 1987, S.185-217

Rothe, Ursula: Distribution of spelling errors by Japanese English-users. IN: ROTHE 1991, S. 168-171

Schwibbe, Gudrun; Schwibbe, Michael H.: Validierungsuntersuchungen zum Duisburg English Language Test for Advanced Students (DELTA). IN: GROTJAHN 1987, S.255-274

Shimamura, Naomi: Kanji no kaitokuritsu - haitō kanji i yoru chigai -. [Acquisition rate of Kanji by school children. The difference due to Kanji assigned to each grade]. IN: MLJAP 17(1990)6, S. 273-279

DIALECTOLOGY

Embleton, Sheila M.: A new technique for dialectometry. IN: Becker, Valerie (Ed.): Twelfth LACUS Forum. Lake Bluff, Ill.: Jupiter Press, 1987. S. 91-98

Embleton, Sheila M.: Multidimensional scaling as a dialectometrical technique. IN: BABITCH 1987, S. 33-49

Goebl, Hans: Considerazioni dialometriche sul problema dell' unità retoromanza (ladinà). IN: MONLAD 12(1988), S. 39-59

Goebl, Hans: Il posto dialettometrico che spetta ai punti-AIS 338 (Adorgnano, Friuli), 398 (Dignano/Vodnjan, Istria), e 367 (Grado, Friuli). IN: LCS 28(1988), S. 75-103

Goebl, Hans: Points chauds de l'analyse dialectométrique. Pondération et visualisation. IN: RLR 51(1987), S. 63-118

Hummel, Lutz: Dialektometrische Analysen zum 'Kleinen Deutschen Sprachatlas (KDSA)'. Experimentelle Untersuchungen zu taxometrischen Ordnungsstrukturen als dialektaler Gliederung des deutschen Sprachraums. Tübingen: Niemeyer, 1990

Inoue, Fumio: New dialect and standard language. Style-shift in Tokyo. IN: ACS 42(1991), S.[49]-68

Kelle, Bernhard: The automatic computation of linguistic maps with the aid of cluster analysis. IN: THOMAS 1987, S. 585-599

Klemola, Juhani: Dialect areas in the South-West of England. An exercise in cluster analysis. Ms., University of Joensuu, Finland. Presented at the International Congress of Dialectologists, Bamberg, July 31, 1990. 1990

Kretzschmar, William A.: Computers and the American Linguistic Atlas. IN: THOMAS 1087, S. 200-224

Linn, Michael D.; Regal, Ronald R.: Verb analysis of the Linguistic Atlas of the North Central States. A case study in preliminary analysis of a large data set. IN: THOMAS 1987, S. 138-154

Lu, Zhiji: A quantitative method of dialect subgrouping. The case of dialects in Jiangsu and Shanghai. IN: LASCI 9(1987), S. 217-229

Veith, Werner H.: Linguistic atlases of German. A study of computer-aided projects. IN: THOMAS 1987, S. 551-556

Viereck, Wolfgang; Ramisch, Heinrich; Händler, Harald; Hoffmann, Petra; Putschke, Wolfgang: The Computer Developed Linguistic Atlas of England (CLAE). Tübingen: Niemeyer, 1990

Viereck, Wolfgang: The computerisation and quantification of linguistic data. Dialectometrical methods. IN: THOMAS 1987, S. 525-550

GRAMMAR

Altmann, Gabriel: Hypotheses about compounds. IN: GLOTT 10(1989), S.[100]-107

Altmann, Gabriel: Word class diversification of Arabic verbal roots. IN: ROTHE 1991, S. 57-59

Best, Karl-Heinz: "Von": Zur Diversifikation einer Partikel des Deutschen. IN: ROTHE 1991, S. 94-104

Best, Karl-Heinz: Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen. IN: GLOTT 11(1990), S.[107]-110

Fuchs, Rinje: Diversifikation der Präposition "auf". IN: ROTHE 1991, S. 105-115

Hammerl, Rolf; Sambor, Jadwiga: Untersuchungen zur Verteilung der Bedeutungen der polyfunktionalen polnischen Präposition "w" im Text. IN: ROTHE 1991, S. 127-137

Hasumi, Yōko: Dōitsu jōhō ni motozuku bunshō hyōgen no ido ni tsuite no bunsetsu. [Difference of expressions on the same information]. IN: MLJAP 18(1991)3, S. 136-144

Hennern, Anja: Zur semantischen Diversifikation von "in" im Englischen. IN: ROTHE 1991, S. 116-126

Junger, Judith: Diversification in the modern Hebrew verbal system. IN: GLOTT 10(1989), S.[71]-99

Kuße, Holger: "A" and "no" in N. M. Karamzins Pis'ma russkogo putešestvennika. IN: ROTHE 1991, S. 173-182

Lichtenberk, Frantisek: On the gradualness of grammaticalization. IN: Traugott, Elizabeth Closs; Heine, Bernd (Eds): Approaches to grammaticalization. 1: Focus on theoretical and methodological issues. Amsterdam u.a.: J. Benjamins publ. comp., 1991. (Typological studies in language ; 19.1) S.37-80

Mergenthaler, Erhard; Pokorny, D.: Die Wortartenverteilung. Eine linguostatistische Textanalyse. IN: Faulbaum; Haux; Jöckel (Eds): Softstat '89. Fortschritte der Statistik Software 2. Stuttgart: Fischer, 1990. S. 512-521

Nemcová, Emília: Semantic diversification of Slovak verbal prefixes. IN: ROTHE 1991, S. 67-74

Raether, Anette; Rothe, Ursula: Diversifikation der deutschen Komposita: "Substantiv plus Substantiv". IN: ROTHE 1991, S. 85-91

Roos, Undine: Zur Diversifikation der japanischen Postposition "ni". IN: ROTHE 1991, S. 75-82

Roos, Undine: Zur Verteilung japanischer wort- bzw. satzsyntaktischer Postpositionen in einem vollendeten Text. IN: GLOTT 12(1990), S.85-92

Rothe, Ursula: Assoziationen und Dissoziationen zwischen dem Genus und der phonetischen Struktur der einsilbigen deutschen Nomina. IN: GLOTT 12(1990), S.93-105

Rothe, Ursula: Polylexy and compounding. IN: GLOTT 9(1988), S.121-134

Rothe, Ursula: Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina. IN: GLOTT 11(1990), S.[111]-121

Rothe, Ursula: The diversification of the case Genitive. IN: ROTHE 1991, S. 140-156

Rothe, Ursula: Verteilung der Suffixe denominaler Veben nach ihren semantischen Wortbildungsmustern. IN: GLOTT 12(1990), S.107-114

Tokunaga, Takenobu; Tanaka, Hozumi: Ketsugōka jōhō ni motozuku nihongo gojun no suitei. [On estimating Japanese word order based on valency information]. IN: MLJAP 18(1991)2, S. 53-65

GRAPHEMICS

Ishii, Hisao: Kana oyobi on no shutsugen hindo no shochōsa. [On the inquiries into the frequency of Kanas and of sounds in modern Japanese]. IN: MLJAP 18(1991)2, S. 84-97

Ishii, Hisao: Otogizōshi "Sagoromo no Chūjō" no kana no shutsugen hindo no kansoku gosa ni tsuite. [Observational error of frequency in language. A case of personal equation]. IN: MLJAP 17(1990)7, S. 328-353

Umeda, Michio: Katachi no ruijijō ni chakumoku shita tango kōsei moji no teiryōteki bunsetsu. [Quantitative analysis of characters in a word based on the similarity of pattern shapes]. IN: MLJAP 17(1990)4, S. 147-168

Yoshida, Masakazu; Seki, Yōko; Koshiba, Ryōsuke: Shomei no haiji sokuteihō - 'konpyûta' ni yoru hisseki sokutei I. [Measuring method of autographic alignment. Measuring of handwriting with a computer I]. IN: MLJAP 17(1990)6, S. 301-308

HISTORICAL LINGUISTICS

Embleton, Sheila M.: Mathematical methods of genetic classification. IN: Lamb, Sydney (Ed.) [u.a.]: Sprung from some common source. Investigations into the prehistory of languages. Stanford, Calif.: Stanford Univ. Press, 1991

INFORMATION THEORY

Guy, Jacques B. M.: Fast high-order monkeys and a fast algorithm for calculating high-order character entropies. IN: GLOTT 12(1990), S.125-130

LEXICOLOGY AND LEXICOGRAPHY

Hammerl, Rolf: Länge-Frequenz, Länge-Rangnummer. Überprüfung von zwei lexikalischen Modellen. IN: GLOTT 12(1990), S.1-24

Hammerl, Rolf: Untersuchungen zur Struktur der Lexik. Aufbau eines lexikalischen Basismodells. Trier: Wissenschaftl. Verl., 1991 ISBN: 3-88476-005-X

Hammerl, Rolf: Untersuchungen zur Verteilung der Wortarten im Text. IN: GLOTT 11(1990), S.[142]-156

Hammerl, Rolf; Sambor, Jadwiga: Vergleich der Längenverteilungen von Lexemen nach der Silbenzahl. Im Lexikon und im Textwörterbuch. IN: GLOTT 10(1989), S.[198]-204

Hammerl, Rolf: Zum Aufbau eines dynamischen Lexikmodells. Dynamische Mikro- und Makroprozesse der Lexik. IN: GLOTT 11(1990), S.[19]-40

Monsell, S.; Doyle, M. C.; Haggard, P. N.: Effects of frequency on visual word recognition tasks. Where are they? IN: JEXPG 118(1989)1, S.43-71

Sambor, Jadwiga: Polnische Version des Projekts "Sprachliche Synergetik. Teil 1. Quantitative Lexikologie". IN: GLOTT 10(1989), S.[171]-197

Schierholz, Stefan: Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive. Tübingen: Niemeyer, 1991 ISBN: 3-484-30269-0

Schweiger, F.: On Dixon's model of lexical diffusion in Australia. IN: GLOTT 11(1990), S.[57]-67

Stellingsma, Hans: Quantitative and qualitative research based on the linguistic database of Frisian. IN: GLOTT 8(1987), S. 192-199

Tuttle, Laurence Heath; Stanish, William: Usage trends in the vocabulary of François Rabelais. IN: GLOTT 9(1988), S.201-218

Zörnig, Peter; Köhler, Reinhard; Brinkmöller, R.: Differential equation models for the oscillation of the word length as a function of frequency. IN: GLOTT 12(1990), S.25-40

METHODOLOGY

Altmann, Gabriel; Hammerl, Rolf: Diskrete Wahrscheinlichkeitsverteilungen 1. Bochum: Brockmeyer, 1989 (Quantitative linguistics ; 41) ISBN: 3-88339-764-4

Grotjahn, Rüdiger: Butler, Christopher: Statistics in linguistics. Oxford, 1985. IN: GLOTT 9(1988), S.247-259 [Review]

Mańczak, Witold: La classification des langues romanes. Kraków: Universitas, 1991 ISBN: 83-7052-032-4

Tambovcev, Jurij A.: Kompaktnost' finno-ugorskoj jazykovoj semji po dannym konsonantnogo koėfficienta. IN: LIUR 26(1990), S. 13-20

MORPHOLOGY

Hammerl, Rolf: Überprüfung einer Hypothese zur Kompositabildung. An polnischem Sprachmaterial. IN: GLOTT 12(1990), S.73-83

Ishii, Hisao: Zasshi ni okeru go no nagasa. [Length of words used in a magazine]. IN: MLJAP 17(1990)4, S. 193-208

Miyajima, Tatsuo: Tango no shiyō dosū to nagasa, furusa. [The relations among the length, oldness and frequency of classical Japanese words]. IN: MLJAP 17(1990)6, S. 287-300

MUSICOLOGY

Agmon, Eytan: Linear transformations between cyclically generated chords. IN: MUSI 3(1991), S.15-40

Bachmutova, I. V.; Gusev, V. D.; Titkova, T. N.: Repetition and variation in melody. Towards a quantitative study. IN: MUSI 2(1990), S. 143-168

Balaban, Mira: Music structures. A temporal hierarchical representation for music. IN: MUSI 2(1990), S. 1-51

Baroni, M.; Callegari, L.: Analysis of a repertoire of eighteenth-century French chansons. IN: MUSI 2(1990), S. 197-240

Bernard, Jonathan W.: Premises and applications of spatial analysis. IN: MUSI 2(1990), S. 241-277

Boroda, Mojsej G.: Rythmical repetition, rythmical variation in Folk music and composed music. Towards a quantitative study. IN: MUSI 2(1990), S. 121-140

Boroda, Mojsej G.: The organization of repetitions in the musical composition. Towards a quantitative-systemic approach. IN: MUSI 2(1990), S. 53-105

Boroda, Mojsej G.; Orlov, Jurij Konstantinovič: Zum Vergleich der Wahrnehmungsgeschwindigkeit von Informationseinheiten in Musik und Rede. IN: MUSI 2(1990), S. 279-283

Boroda, Mojsej Grigor'evič; Altmann, Gabriel: Menzerath's Law in musical texts. IN: MUSI 3(1991), S.1-13

Boroda, Mojsej Grigor'evič: Ritmičeskie modeli v fol'klornoj melodike. K probleme kvantitativnogo analiza. [Rythmical models in folk tunes. Towards the quantitative approach]. IN: Alekseev, E.; Andreeva, E.; Boroda, M. G.; Tangjan, A. (Eds): Količestvennye metody v muzykal'noj fol'kloristike i muzykoznanii. Moskva: Sovetskij kompositor. S. 36-85

Boroda, Mojsej Grigor'evič: Rythmic models in music: towards the quantitative study. IN: MUSI 3(1991), S.123-162

Boroda, Mojsej Grigor'evič: The concept of "metrical force" in music with bar structure. IN: MUSI 3(1991), S. 59-94

Boroda, Mojsej Grigor'evič: Towards the basic semantic units of a musical text. IN: MUSI 1(1988), S.11-67

Boroda, Mojsej Grigor'evič (Ed.): Musikometrika 1. Bochum: Brockmeyer, 1988 (Quantitative linguistics ; 37) ISBN: 3-88339-678-8

Boroda, Mojsej Grigor'evič (Ed.): Musikometrika 2. Bochum: Brockmeyer, 1990 (Quantitative linguistics ; 43) ISBN: 3-88339-793-8

Boroda, Mojsej Grigor'evič (Ed.): Musikometrika 3. Bochum: Brockmeyer, 1991 (Quantitative linguistics ; 46) ISBN: 3-88339-905-1

Detlovs, Vilnis K.: Modal functions and intervalic structure of melody. IN: MUSI 3(1991), S.41-58

Halperin, D.: A segmentation algorithm and its application to medieval monophonic music. IN: MUSI 2(1990), S. 107-119

Lippus, Urve: On the possibility of historical study of Estonian traditional melodies. IN: MUSI 2(1990), S. 187-195

Rüütel, I.; Haugas, K.: A method for distinguishing melody types and establishing typological groups. On the material of Estonian runo songs. IN: MUSI 2(1990), S. 169-186

Rüütel, I.; Haugas, K.: Metod raspoznavanija melodičeskich tipov i opredelenija tipologičeskich grupp. IN: Alekseev, E. E. (Ed.) ; Andreeva, E. D. (Ed.) ; Boroda, M. G. (Ed.) ; Tangjan, A. S. (Ed.): Količestvennye metody v muzykal'noj fol'kloristike i muzykoznanii. Moskva: Sovetskij Kompositor, 1988.

Tanguiane, Andranick S.: Recognition of chords, perception correlativity, and music theory. IN: MUSI 3(1991), S.163-199

PHONETICS

Cichocki, Wladyslaw: Uses of dual scaling in social dialectology. Multidimensional analysis of vowel variation. IN: THOMAS 1987, S. 187-199

Ljublinskaja, Valentina; Sappok, Christian: Die Beurteilung der Höhe von Melodiekonturen beim Sprachsignal. IN: MUSI 3(1991), S.201-217

PHONOLOGY

a Campo, Frank W.; Geršić, Slavko; Naumann, Carl L.; Altmann, Gabriel: Subjektive Ähnlichkeit deutscher Laute. IN: GLOTT 10(1989), S.[46]-70

Abu-Salim, Issam; Abd-El-Jawad, Hassan: Syllable patterns in Levantine Arabic. IN: STLSCI 18(1988), S. 1-22

Altmann, Gabriel: Tendenzielle Vokalharmonie. IN: GLOTT 8(1987), S. 104-112

Baevskij, V. S.; Osipova, L. Ja.: The algorithm and some results of a statistical investigation of alternating rhythm on the Minsk-32 computer. IN: GLOTT 8(1987), S. 157-177

Enninger, Werner; Köhler, Reinhard; Korda, Helge; Raith, Joachim: Zur Sprachunabhängigkeit suprasegmentaler phonetischer Eigenschaften. IN: GLOTT 8(1987), S. 57-103

Geršić, Slavko; Altmann, Gabriel: Ein Modell für die Variabilität der Vokaldauer. IN: GLOTT 9(1988), S.49-58

Guiter, Henri: Arborescences linguistiques. IN: GLOTT 9(1988), S.171-200

Izumi, Asako; Mizutani, Shizuo: Tsutsui Yasutaka "Zanzō ni Kuchibnei o" no onbunpu hoi. [Supplement to 'On sound-distribution in Tutui Yasutaka's "Zanzo ni Kutibeni wo"']. IN: MLJAP 18(1991)2, S. 80-83

Rothe, Ursula; Zörnig, Peter: The entropy of phoneme frequencies. German and French. IN: GLOTT 11(1990), S.[198]-205

Schulz, Klaus-Peter; Altmann, Gabriel: Lautliche Strukturierung von Spracheinheiten. IN: GLOTT 9(1988), S.1-47

Tambovcev, Jurij A.: Selected phonostatistical features of the Khakas language. IN: UAJB 63(1991), S. 155-165

PSYCHOLINGUISTICS

Dolinskij, V. A.: Raspredelenie reakcij v èxperimentach po verbal'nym associacijam. IN: Acta et commentationes universitatis Tartuensis 827(1988), S. 89-101

Frauenfelder, U. H.: Structure and computation in the human mental lexicon. IN: Haken, Hermann ; Stadler, Michael (Eds): Synergetics of cognition. Proceedings of the international symposium at Schloß Elmau, Bavaria, June 4-8, 1989. Berlin u. a.: Springer, 1990. (Springer series in synergetics ; 45). S.406-414

Givón, T.: Serial verbs and the mental reality of "event". Grammatical versus cognitive packaging. IN: Traugott, Elizabeth Closs ; Heine, Bernd (Eds): Approaches to grammaticalization. 1: Focus on theoretical and methodological issues. Amsterdam u. a.: J. Benjamins publ. comp., 1991. (Typological studies in language ; 19.1). S.81-128

Grotjahn, Rüdiger: Nowakowska, Maria: Quantitative psychology. Some chosen problems and new ideas. Amsterdam, 1983. IN: GLOTT 9(1988), S.235-245 [Review]

Monsell, S.; Doyle, M. C.; Haggard, P. N.: Effects of frequency on visual word recognition tasks. Where are they? IN: JEXPG 118(1989)1, S.43-71

SEMANTICS

Altmann, Gabriel; Best, Karl-Heinz; Kind, B.: Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. IN: GLOTT 8(1987), S. 130-139

Hammerl, Rolf: Neue modelltheoretische Untersuchungen im Zusammenhang mit dem Martingesetz der Abstraktionsebenen. IN: GLOTT 9(1988), S.105-119

Hammerl, Rolf: Untersuchung struktureller Eigenschaften von Begriffsnetzen. IN: GLOTT 10(1989), S.[141]-154

Kimura, Mutsuko; Ogino, Takano; Kinukawa, Hiroshi; Kurabayashi, Yasuko: A study on the possibility of discriminating between homonyms by a semantic correlation approach considering their preceding and succeeding words. IN: GLOTT 11(1990), S.[122]-141

Rieger, Burghard B.: Unscharfe Semantik. Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten. Frankfurt am Main [u.a.]: Lang, 1989 ISBN: 3-631-41704-7

Rothe, Ursula: Semantische Motivation der Genuszuweisung. IN: GLOTT 11(1990), S.[95]-106

Takeuchi, Haruhiko: 'Faji' hyōteihō ni yoru teido hyōgen yōgo no imi keisoku. [Measuring the meaning of linguistic hedges by fuzzy rating scale method]. IN: MLJAP 17(1991)8, S.365-376

SYNTAX

Altmann, Gabriel: Verteilungen der Satzlängen. IN: GLOTT 9(1988), S.147-169

Nakano, Hiroshi: '(te) itadaku' bun ni okeru shōryaku. [Ellipsis of the agent and the patient in the "te itadaku" sentence]. IN: MLJAP 18(1991)2, S. 66-79

Schweers, Anja; Jinyang, Zhu: Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. IN: ROTHE 1991, S. 157-165

Tanaka, Akio: Stochastic model of sentence structure in Japanese literature. IN: GLOTT 11(1990), S.[172]-197

TEXTOLOGY

Altmann, Gabriel: Wiederholungen in Texten. Bochum: Brockmeyer, 1988 (Quantitative linguistics ; 36) ISBN: 3-88339-663-X

Bachmutova, I. V.; Gusev, V. D.; Titkova, T. N.: Zakonomernosti var'irovanija v tekstach različnoj prirody i technika ich kvantitativnogo issledovanija. [The regularities of the variation in texts of different origin and technique of their quantitative study]. IN: VCS 123(1987), S.25-49

Beliaeva, I.; Shingariova, E.: Ilpo Tapani Piirainen and Mariann Skog-Södersved: Untersuchungen zur Sprache der Leitartikel in der "Frankfurter Allgemeinen Zeitung". Vaasa: Vaasan Korkeakoulun Julkaisuja, 1982. (Tutkimuksia ; 84) (Philologie ; 10), 95 pp.. IN: GLOTT 8(1987), S. 189-191 [Review]

Boroda, Mojsej Grigor'evič; Dolinskij, V. A.: Problems of quantitative text analysis. A survey of seminars held by the research group "Text as an object of interdisciplinary investigations" in Voronovo (USSR), 8-12 October 1985 and Tartu (USSR), 27-31 January 1986). IN: GLOTT 9(1988), S.135-145

Boroda, Mojsej Grigor'evič; Polikarpov, A. A.: The Zipf-Mandelbrot law and units of different text levels. IN: MUSI 1(1988), S.127-158

Ermolenko, G. V.: Anonimnye proizvedenija i ich avtory. Na materiale russkich tekstov vtoroj poloviny 19 - načala 20 v. [Anonyme Werke und ihre Autoren. Am Material russischer Text von der zweiten Hälfte des 19. bis zum Beginn des 20. Jh.]. Minsk: Izd. universitetskoe, 1988 ISBN: 5-7855-0007-8

Hřebíček, Luděk: A syntactic variable on the text level. IN: GLOTT 10(1989), S.[205]-218

Krasnoperova, M. A.: The relationships between degrees of contrast in rhythmic structures. IN: GLOTT 8(1987), S. 140-156

Malmanger, Curtis Alvin: A statistical word study of the Book of Hebrews as to its Pauline authorship. 1988 zugl.: Diss., Univ. of Northern Colorado, 1988

Rumpel, Dieter: On the internal structure of the diskos of Phaistos text. IN: GLOTT 12(1990), S.131-149

Teufel, Bernd: Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natürlichsprachigen Texten. Zürich: Verl. d. Fachvereine, 1989 (Informatik-Dissertationen ETH Zürich ; 13)

Trauth, Michael: The Phaistos disc and the devil's advocate. On the apories of ancient research. IN: GLOTT 12(1990), S.151-173

Vasjutočkin, G. S.: Das rhythmische System der "Alexandrinischen Gesänge". IN: GLOTT 8(1987), S. 178-188

Zörnig, Peter: A theory of distances between like elements in a sequence. IN: GLOTT 8(1987), S. 1-22

THEORY

Altmann, Gabriel: Modelling diversification phenomena in language. IN: ROTHE 1991, S. 33-46

Altmann, Gabriel; Best, Karl-Heinz; Kind, B.: Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. IN: GLOTT 8(1987), S. 130-139

Altmann, Gabriel; Schwibbe, Michael H.: Das Menzerath'sche Gesetz in informationsverarbeitenden Systemen. Hildesheim: Olms, 1989

Beöthy, Ersébet; Altmann, Gabriel: The diversification of meaning of Hungarian verbal prefixes. 1. "-meg". IN: ROTHE 1991, S. 60-66

Best, Karl-Heinz; Beöthy, Ersébet; Altmann, Gabriel: Ein methodischer Beitrag zum Piotrowski-Gesetz. IN: GLOTT 12(1990), S.115-124

Hammerl, Rolf; Maj, Jaroslaw: Ein Beitrag zu Köhlers Modell der sprachlichen Selbstregulation. IN: GLOTT 10(1989), S.[1]-31

Hammerl, Rolf: Neue Perspektiven der sprachlichen Synergetik. Begriffsstrukturen, kognitive Gesetze. IN: GLOTT 10(1989), S.[129]-140

Hammerl, Rolf: Untersuchungen zur mathematischen Beschreibung des Martingesetzes der Abstraktionsebenen. IN: GLOTT 8(1987), S. 113-129

Hřebíček, Luděk: Menzerath-Altmann's law on the semantic level. IN: GLOTT 11(1990), S.[47]-56

Hřebíček, Luděk: The constants of the Menzerath-Altmann law. IN: GLOTT 12(1990), S.61-71

Köhler, Reinhard: Diversification of coding methods in grammar. IN: ROTHE 1991, S. 47-55

Köhler, Reinhard: Elemente der synergetischen Linguistik. IN: GLOTT 12(1990), S.179-187

Köhler, Reinhard: Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. IN: GLOTT 11(1990), S.[1]-18

Köhler, Reinhard: Systems theoretical linguistics. Offprint [von TL, ersch. 1989] Berlin [u.a.]: deGruyter, IN: TL 14(1987), S. 241-257

Köhler, Reinhard: Zur Charakteristik dynamischer Modelle. Anmerkungen zu einem Beitrag von R. Hammerl und J. Maj. IN: GLOTT 11(1990), S.[41]-46

Maj, Jaroslaw: Kybernetische Aspekte des synergetischen Modells von R. Köhler. IN: GLOTT 12(1990), S.175-177

Rothe, Ursula: Diversification processes in grammar. An introduction. IN: ROTHE 1991, S. 3-32

Schierholz, Stefan: Kritische Aspekte zum Martinschen Gesetz. IN: GLOTT 10(1989), S.[108]-128

Wildgen, Wolfgang: Basic principles of self-organization in language. IN: Haken, Hermann ; Stadler, Michael (Eds): Synergetics of cognition. Proceedings of the international symposium at Schloß Elmau, Bavaria, June 4-8, 1989. Berlin u. a.: Springer, 1990. (Springer series in synergetics ; 45). S.415-426

Wildgen, Wolfgang: Bio- und neurolinguistische Aspekte der dynamischen Sprachtheorie. IN: GLOTT 8(1987), S. 23-56

Wildgen, Wolfgang; Mottron, Laurent: Dynamische Sprachtheorie. Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese. Bochum: Brockmeyer, 1987 (Quantitative linguistics ; 33) ISBN: 3-88339-619-2

Wildgen, Wolfgang: L'instabilité du langage et sa capacité d'auto-organisation. IN: RSSI 9(1989)1-3, S.53-80