[*Mechanical Translation*, vol.1, no.3, December 1954; pp. 38-40]

THE DISTRIBUTION OF WORD LENGTH IN TECHNICAL RUSSIAN

Anthony G. Oettinger
Computation Laboratory, Harvard University

IN the course of an analysis of several samples of technical Russian undertaken as part of a study in mechanical translation, a number of statistical data reflecting the structure of these samples were compiled. One of these, the distribution of word length, is presented here as Fig. 1.

The theoretical interest of this distribution arises from the possibility of using it as a basis for an operational definition of words in printed texts. If texts are considered purely as sequences of symbols including the letters, punctuation marks, and space, the resulting sequences are of a length which no practicable machine can manage. A study of the distribution of the number of symbols between pairs of successive symbols of certain classes would be one way to reveal structural characteristics of the text sequences potentially useful toward the definition of manageable and significant subsequences. The subsequences included between successive occurrences of letter pairs have not been investigated. Those included between successive pairs of periods, exclamation points or question marks can be identified with the classical sentence, and finally, those included between successive pairs of punctuation marks or spaces can be identified with words. The length distribution of the latter subsequences has the desirable property, not shared by the others, of being concentrated at relatively low values of length, and of having no elements exceeding a certain length (Fig. 1). Words, defined in this fashion, can readily be identified by a machine and they are of limited variety, so that their listing in a dictionary is practicable.

From the practical point of view, the distribution is useful in planning input and storage facilities in experimental translating equipment.

The samples used were relatively small, and Fig. 1 should therefore be interpreted with great caution. The bar graph represents the distribution of a sample totalling 6,486 words. Points are used to indicate the distributions obtained from smaller constituents of the total. The scattering is such as to indicate that samples 1, 2, and 3 differ significantly among each other in details of their distributions. An examination of the texts indicates that these differences can safely be attributed to differing subject matter and styles. However, all distributions are bimodal, perhaps trimodal, and cut off at k=18. The mode about k= 7 is attributable to the large number of different words used to define the particular subject of each text. The peaks at k= 1 and at k= 3 are due to a small number of very frequent "grammatical words," that is, prepositions, conjunctions, etc. The five most frequent words of length 1, 2, and 3 in the total sample are listed in Table 1. This table shows that the most frequent two letter words are consistently less frequent than three letter words of similar rank. One and two letter words are exclusively grammatical; 90% of the three letter words are also grammatical, leaving 10% dependent on the subject matter. The words of length 4 are nearly all inflected. The fact that only very few Russian words have stems of three or less letters probably accounts for the valley at k= 4. Indications thus are that the modal and cut-off structure of the distributions are functions of the structure of the Russian language, while variations within these structures are characteristic of individual authors. For those who might wish to draw their own conclusions, the raw data is given in Table 2, and the sources of the samples are listed in Table 3. Letter, diagram and suffix distributions compiled from the same samples may be found in the reference.

TABLE 1

v	210	na	86	pri	93
i	165	iz	57	dlja	72
s	91	po	46	chto	50
k	43	ot	28	kak	29
a	21	ne	26	ili	22

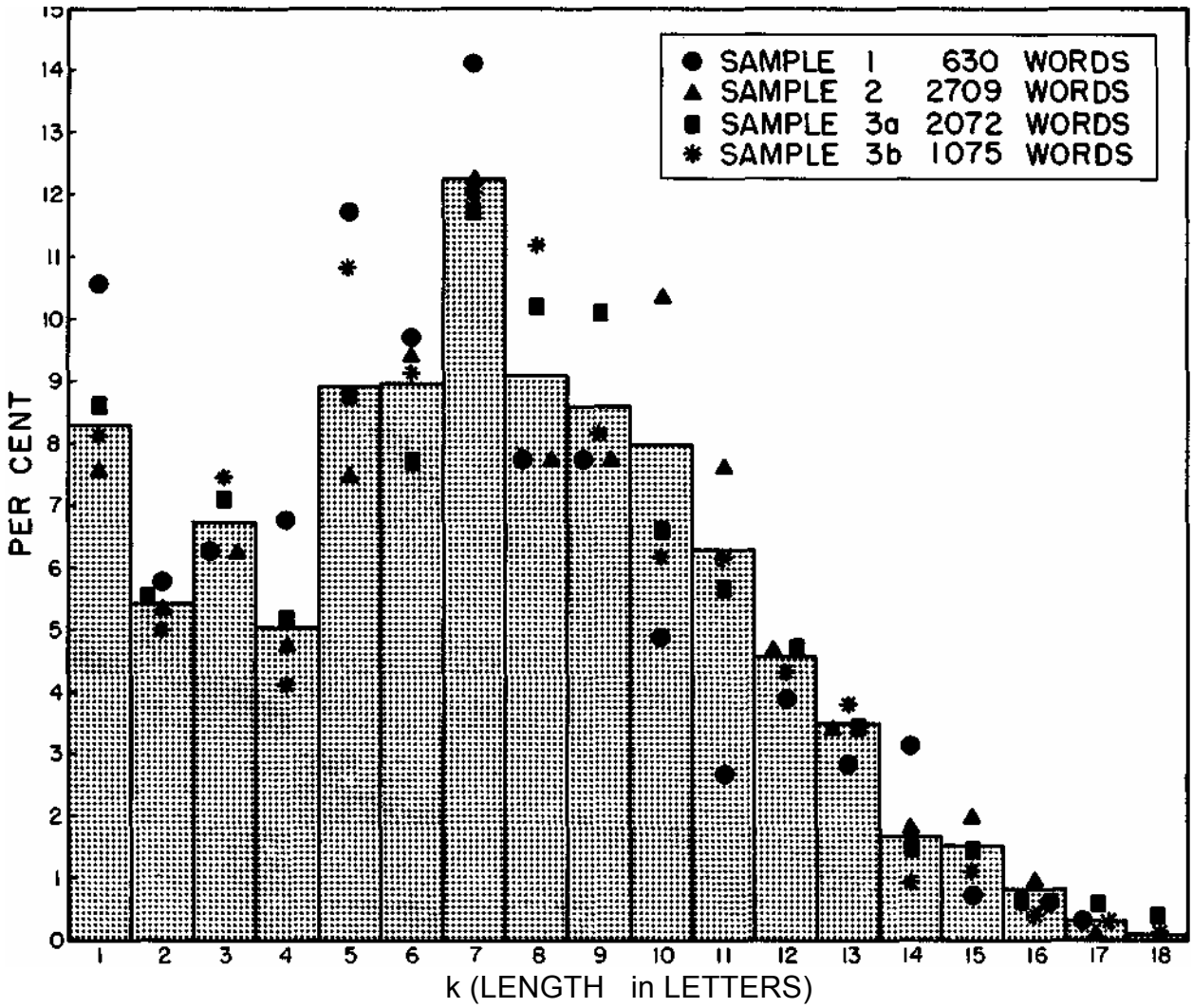
SAMPLE 1 630 WORDS
SAMPLE 2 2709 WORDS
SAMPLE 3a 2072 WORDS
SAMPLE 3b 1075 WORDS
PER CENT
k (LENGTH in LETTERS)

Figure 1

TABLE 2

Word length	Frequency				
	Sample 1	Sample 2	Sample 3a	Sample 3b	Total
1	67	204	178	88	537
2	36	147	114	54	351
3	40	170	148	80	438
4	43	130	107	45	325
5	74	203	183	117	577
6	61	258	161	99	579
7	89	332	245	129	795
8	49	209	212	121	591
9	49	209	211	88	557
10	31	281	138	67	517
11	17	208	118	66	409
12	25	127	98	47	297
13	18	94	72	41	225
14	20	50	29	10	109
15	5	54	28	13	100
16	4	28	16	5	53
17	2	5	9	4	20
18	0	0	5	1	6

TABLE 3

1. A. G Lunts, 1950, "Prilozhenie Matrichnoj Bulevskoj Algebry k Analizu i Sintezu Relejno-Kontaktnyx Sxem," Doklady Akademii Nauk SSSR, 70, pp. 421-23.

2. K. V. Valdimirskij, 1951, "O Sinxronnom Fil'tre," Zhurnal Eksperimental'noj i Teoreticheskoj Fiziki, *21,* pp. 2-10.

3. B. P. Aseev, 1947, Osnovy Padiotexniki (Moskva: Svjaz'izdat) (a) pp. 10, 18, 20, 21, 23, 33, 37, 42, 45, 49, 55 (part); (b) pp. 55 (part), 59, 64, 65, 71, 122

REFERENCE

Oettinger, A. G., "A Study for the Design of an Automatic Dictionary," Doctoral Thesis, Harvard University (1954).