A MATHEMATICAL MODEL OF THE VOCABULARY-TEXT RELATION

Juhan Tuldava

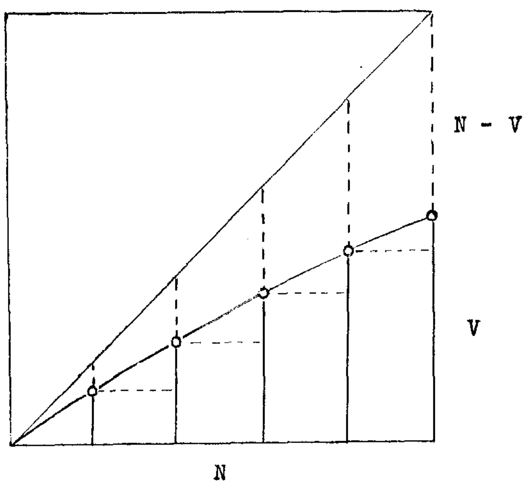
Tartu, Estonia, USSR

A new method for calculating vocabulary size as a function of text length is discussed. The vocabulary growth is treated as a probabilistic process governed by the principle of "the restriction of variety" of lexics. Proceeding from the basic model of the vocabulary-text relation a formula with good descriptive power is constructed. The statistical fit and the possibilities of extrapolation beyond the limits of observable data are illustrated on the material of several languages belonging to different typological groups.

1. There are a great number of attempts to construct an appropriate mathematical model which would express the dependence of the size of the size of vocabulary (V) on the size of text (N). This is not only of practical importance for the resolution of a series of problems in the automatic processing of texts, but it is also connected with the theoretical explanation of some important aspects of text generation. In practice one often makes use of various empirical formulae which describe the growth of vocabulary with sufficient precision in the case of concrete texts and languages[1], though such formulae do not have any general significance. Of special interest are some "complex" models derived from theoretical considerations, e.g., by basing one's considerations on the hypothesis about the lognormal distribution of words in a text (Carroll)[2] or by deducing the relation between V and N from some other important quantitative characteristics of text such as Zipf's law and Yule's distribution (Kalinin, Orlov)[3]. The author underlines the importance of these conceptions for the theory of quantitative linguistics on the whole, but points out their insufficiency in solving some practical linguo-statistical problems where greater exactness and reliability are needed (stylo-statistical analysis, text attribution, extrapolation beyond the limits of observable data, etc.).

2. Instead of the "complex" models a "direct" method is proposed where the relation between V and N is regarded as the primary component with its own immanent properties in the statistical organization of text. The relation between V and N has to be analyzed on the background of some essential inner factors of text generation. The dynamics of vocabulary growth is considered as the result of the interaction of several linguistic and extra-linguistic factors which in an integral way are governed by the principle of "the restriction of variety" of lexics (an analogue of the principle of the decrease of entropy in self-regulating systems). The concept of the variety of lexics is defined as the relation between the size of vocabulary and the size of text in the form of V/N (type-token ratio, or coefficient of variety) or N/V (average frequency of word occurrences).

The coefficient of variety is supposed to be correlated with the probabilistic process of choosing "new" (unused) and "old" (already used in the text) words at each stage of text generation. The steady decrease of the degree of variety $V/N = p$ is attended by the increase of its counterpart: $(N - V)/N = 1 - V/N = q$ $(p + q = 1)$, which can be interpreted as the "pressure of recurrency" of words in real texts (analogous to the concept of redundancy in the theory of information):



3. The formulae of the relation between V and N are constructed from the basic models: $V = Np$ or $V = N(1 - q)$. For this purpose the quantitative changes of $V/N = p$ depending on the size of text are analyzed. According to the initial hypothesis the relation between V/N and N is approximated by the power function of the type: $V/N = aN^B$ (a and B are constants; $B < 0$), which leads to the well-known formula of G. Herdan[4]: $V = aN^b$ (where $b = B + 1$). A verification shows good agreement with empirical data in the initial stages of text formation (in the limits of about 4,000 - 5,000 tokens which correspond to a short communication). Later on the rate of the diminishing of the degree of variety (V/N) gradually slows down (due to the rise of new themes in the course of text generation). Accordingly the initial formula has to be modified and this can be done by logarithmization of the variables. The first attempt gives us $\ln (V/N) = aN^B$, which leads to some variants of the Weibull distribution. This kind of distribution shows good agreement with the empirical data within the boundaries of a text of medium length, but it is not good for extrapolation. Only after balancing the initial formula by the logarithmization of both variables we obtain $\ln (V/N) = a(\ln N)^B$ and the corresponding formula for expressing the relation between V and N:

$$V = Ne^{-a(\ln N)^B},$$

or $V = N^{1 - a(\ln N)^b}$ (where $b = B - 1$)[5], which turns out to be the most adequate formula for solving our problems. The constants a and B (which, of course, are not identical with those of the previously mentioned formulae) may be determined on the basis of linearization: $\ln\ln (N/V) = A + B \ln\ln N$, where $A = \ln a$, using the method of least squares. In principle it would be sufficient to have two empirical points for the calculation of the values of the constants but for greater reliability more points are needed.

4. The good descriptive power of the given function and the possibili-

ties of extrapolation in both directions (from the beginning up to a text of about $N = 10^7$) has been verified on the basis of experimental material taken from several languages belonging to different typological groups (Estonian, Kazakh, Latvian, Russian, Polish, Czech, Rumanian, English). The function may be applied to the analysis of individual texts as well as composite homogeneous (similar) texts and the size of vocabulary (V) may be determined by counting either word forms of lexemes. (See Tables 1 and 2.) This seems to corroborate the assumption about the existence of a universal law (presumably of phylogenetic origin) which governs the process of text formation on the quantitative level.

Table 1

The empirical size (V) and the teoretical size (V′) of vocabulary plotted against the length of the text (N). The formula:

$$V' = Ne^{-a(\ln N)^B}$$

a) Latvian newspapers (lexemes)[6]

N	V	V′
50000	7065	7025
100000	9834	9919
200000	13389	13510
300000	16103	15912
10^6	-	24000
10^7	-	37000

(a = 0.003736, B = 2.6304)

b) Czech texts of technical sciences (word forms)[7]

N	V	V′
25000	4829	4827
75000	9603	9626
125000	13056	13050
175000	15858	15853
10^6	-	40000
10^7	-	114000

(a = 0.01123, B = 2.1539)

c) Kazakh newspapers (word forms)[8]

N	V	V′
25000	9088	9161
50000	15047	14875
100000	23895	23523
150000	29785	30378
10^6	-	87000
10^7	-	230000

(a = 0.001372, B = 2.8488)

d) Polish belles-lettres (word forms)[9]

N	V	V′
12172	3434	3458
29787	6146	6044
48255	8026	7998
64510	9250	9398
10^6	-	33000
10^7	-	60000

(a = 0.00364, B = 2.6081)

e) English texts on electronics (word forms)[10]

N	V	V′
50000	5399	5437
100000	7853	7728
150000	9361	9371
200000	10582	10682
10^6	-	20000
10^7	-	38000

(a = 0.009152, B = 2.3057)

f) Rumanian texts on electronics (word forms)[11]

N	V	V′
50000	6785	6841
100000	10281	10070
150000	12477	12479
200000	14292	14454
10^6	-	30000
10^7	-	68000

(a = 0.008148, B = 2.3086)

g) Russian texts on electronics (word forms)[12]

N	V	V′
50000	9464	9388
100000	14062	14168
150000	17263	17803
200000	21468	20818
10^6	-	45000
10^7	-	94000

(a = 0.004284, B = 2.5058)

Table 2

Prediction on the basis of two empirical points (marked with an asterisk)

a) English: literary texts[13] (word forms)

N	V	V′
10051	3009*	3009
101566	13706*	13709
Prediction:		
10	-	9
100	-	78
1000	-	534
2000	700-1000	917
50721	8749	8905
253538	23655	23447
1014232	50406	49280
10^7	-	140000

(a = 0.007879, B = 2.2652)

b) Estonian: A. H. Tammsaare′s novel "Truth and Justice" I (lexemes)[14]

N	V	V′
10000	2114*	2114
20000	3124*	3124
Prediction:		
114124 (the whole book)	7348	7207

(a = 0.006714, B = 2.4521)

c) Russian: A. S. Pushkin′s "Queen of Spades" (lexemes)[15]

N	V	V′
1000	462*	462
2000	787*	787
Prediction:		
3000	1067	1068
4000	1348	1321
5000	1541	1556
6000	1752	1776
6861 (the whole book)	1928	1957

(a = 0.01699, B = 1.9747)

References

1. Kuraszkiewicz, W., Statystyczne badanie słownictwa polskich tekstów XVI wieku. In: Z Polskich Studiów Slawistycznych. Warszawa 1958.

Guiraud, P., Problèmes et méthodes de la statistique linguistique. Dordrecht 1959.

Somers, H. H., Analyse mathématique de langage. Louvain 1959.

Miller, W., Wortschatzumfang und Textlänge. In: Muttersprache, 81. Jg., Nr. 4. Mannheim-Zürich 1971.

Nešitoj V. V., Dlina teksta i ob′em slovarja. In: Metody izučenija leksiki. Minsk 1975.

2. Carroll, J. B., On Sampling From a Lognormal Model of Frequency Distribution. In: Computational Analysis of Present-Day American English. Providence, R. I., 1967.

3. Kalinin, V. M., Nekotorye statističeskie zakony matematičeskoj lingvistiki. In: Problemy kibernetiki, 11, 1964.

Orlov, Yu. K., Obobščennyj zakon Zipfa-Mandelbrota i častotnye struktury informacionnykh edinic različnykh urovnej. In: Vyčislitelnaja lingvistika. Moscow 1976.

4. Herdan, G., Quantitative Linguistics. London 1964.

5. Tuldava, J., Quantitative Relations between the Size of Text and the Size of Vocabulary. In: Journal of Linguistic Calculus. SMIL Quarterly 1977:4.

6. Latviešu valodas biežuma vārdnīca. II:1. Ed. T. Jakubaite. Riga 1969.

7. Bečka, J. V., La structure lexicale des textes téchniques en tchèque. In: Philologica Pragensia, 1972, vol. 15, No. 1.

8. Akhabayev, A., Statističeskij analiz leksiko-morfologičeskoj struktury jazyka kazakhskoj publicistiki. Alma-Ata 1971.

9. Sambor, J., Analiza stosunku "type-token", czyli objętości (W) i długości tekstu (N). In: Prace filologiczne. Tom XX. Warszawa 1970.

10. Alekseev, P. M., Leksičeskaja i morfologičeskaja statistika anglijskogo pod′jazyka elektroniki. In: Statistika reči. Leningrad 1968.

11. Ežan, L. J., Opyt statističeskogo opisanija naučno-tekhničeskogo stilja rumynskogo jazyka. Leningrad 1966.

12. Kalinina, E. A., Častotnyj slovar′ russkogo pod′jazyka elektroniki. In: Statistika reči. Leningrad 1968.

13. Kučera H., and Francis, W. N. (ed.), Computational Analysis of Present-Day American English. Providence, R. I. 1967.

14. Villup, A., A. H. Tammsaare romaani "Tõde ja õigus" I köite autori- ja tegelaskõne sagedussõnastik. In: Acta et Commentationes Universitatis Tartuensis. Vol. 446. Tartu 1978.

15. Orlov, Yu. K., Model′ častotnoj struktury leksiki. In: Issledovanija v oblasti vyčislitel′noj lingvistiki i lingvostatistiki. Moscow 1978.