M. V. Arapov - M. M. Herz

FREQUENCY AND AGE AS CHARACTERISTICS OF A WORD

1. The problem of relation between the frequency and age of a word is only a small part of the general problem of opposition of the synchronic and diachronic aspects of language. The frequency is obviously a purely synchronic characteristic of a word whereas the age (the time interval t between the appearance of the word and the present moment) is a purely diachronic one. However there is a simple dependence between both characteristics: the old age of a word corresponds to a high frequency ranking and vice-versa: among the words with low frequency the proportion of ancient words is small. The existence of this dependency was first discovered by G. K. Zipf (1947).

2. To obtain this dependency in analytical form let us split the whole frequency dictionary into a number of groups of equal size (n words in each of the groups).[1] Each group consists of words of equal or nearly equal values of frequency. The most frequently used n words belong to the group with rank 1, the following n words constitute the group with rank 2 and the words having in the dictionary numbers from $(i-1)\,n+1$ till $i{\cdot}n$ constitute a group with rank i.

Then the ratio $N(i, t)$ of words with an age equal to or greater than t in the i-th group of the dictionary is:

$$N(i, t) = exp\ (-Kt\sqrt{i}) \tag{1}$$

where K is a constant.

3. There is a simple method to obtain (2) from (1). Formula (2) gives us the distribution in the dictionary of words with age t, $t_1 < t \leq t_2$:

$$\begin{aligned} N(i, t_1, t_2) &= N(i, t_1) - N\ (i, t_2) = \\ &= exp\ (-Kt_1\sqrt{i}) - exp\ (-Kt_2\sqrt{i}) \end{aligned} \tag{2}$$

[1] In our experiments $n = 100$.

The interpretation of (2) as a function of rank i is as follows: the number of words that appeared in the time interval t_1-t_2 (for example: from 973 till 1073 A. D.) is growing till it reaches some rank i_{max}, where the function (2) will obtain its maximum value and then it will decrease approaching zero. The maximum value function (2) is the nearer to the beginning of the rank axis the more ancient the epoch considered. The maximum value for the above cited example places it in the group with rank 16 ($i_{max} = 16$); if we chose the XVII-th century as a time interval the point of maximum value would shift to the group with rank 134 ($i_{max} = 134$). Here we have supposed the mean value of K to be 0.025 which holds for a number of European languages. The value of function (2) at the point of maximum is the greater the more proximate to our day are the time intervals we choose. Here we have a sort of mapping of the time axis on the rank 1.

4. The dependency between age and frequency is a dependency "in the mean". There is no way of predicting, for any word with a given age, the frequency of that particular word. All predictions would be applied only to sets of words. By (1) we can obtain the ratio of words in the considered group i having an age equal to/greater than t, but by no means can we establish exactly which words have such an age.

One of the basic postulates from which (1) was established consists in the following: there exists no difference between new and old elements in their behaviour on each step of vocabulary evolution, i.e. the probability of "dissociation" is equal (or nearly equal) for all the words of the same group.

5. Another basic postulate of our theory is as follows. There is no intrinsic difference between any two vocabularies which can be described by (1) or similar functions if this pair of vocabularies represents two historical stages of the evolution of the same language. In other words the collation of a pair of frequency dictionaries compiled for two successive periods in the history of the same language does not in itself (without consideration of historical evidence) make it possible to determine which of the dictionaries corresponds to the later (resp. previous) epoch.

6. Then it is necessary to assume the independence in frequency shifting of any particular word. This assumption rules out the appli-

cation of our theory in other than the vocabulary levels of the language system, since it excludes the situation when dissociation of a particular element yields a " chain reaction " which results in the whole rebuilding of the system.

7. The following problems would be solved within the framework of the present theory:

a) if we have reason to suppose that a given set S of words appeared in a period preceeding some moment t (the value of parameter K is given), we can calculate the absolute value of t; examples of S: the set of words appeared before the splitting of the parent language; the set of words experienced the influence of a given phonetic law, etc.; this problem is the same as the one which glottochronology has attempted to solve.

b) if we have reason to suppose that a given set S of words appeared in a time period following some moment t, we can obtain the absolute value of t (in this case the words belonging to S are innovations, specifically, borrowings) or demonstrate that this moment does not exist (in this case, the given set S most probably consists of vernacular words).

c) having historical data about the time of appearance of the words forming the initial part of the frequency dictionary (1000-3000 most frequent words) we can calculate, with the aid of (1), indexes of the evolution rate K ($K = K(t)$) for the whole vocabulary and for any stage of the language for which historical data is obtainable.

d) analogously we can study with the aid of (2) the dynamics of language borrowing and obtain indexes of intensity of vocabulary interaction for any period for which corresponding data is available.

8. For testing the present theory we have taken the frequency lists for the following languages: Russian (two lists), Czech (one), French (two), Roumanian (three), Spanish (one), German (two) and English (two). The volume of the lists is spread from 600 to 6000 words. The period for which lexicological data are available (depth of analysis did not exceed 1000-1500 years) was present as a system of intervals: $(0, t_1)$, $(0, t_2)$, $(0, t_3)$..., where $t_1 < t_2 < t_3$ ... and zero corresponds to the present moment. We take t as an age of the word x if x was primarily registered in the language at t_r, $t_r > t$ and there exists no such time interval $(0, t_k)$ that $t_r > t_k > t$.

The observed dependency $N(i, t)$ upon t only roughly, for a long period, corresponds to the theoretical derived one, hence we sought to investigate also fluctuation of vocabulary change rate ($K = K(t)$).

A full account on theoretical and experimental aspects of the problem "age and frequency" is contained in M. V. ARAPOV, M. M. HERZ (1972).

REFERENCES

M. V. Arapov, M. M. Herz, *Izmenenie slovarja vo vremeni*, in « Informatsionnye voprosy semiotiki, lingvistiki i avtomatitsheskogo perevoda», III (1972).

G. K. Zipf, *Prehistoric 'cultural strata' in the evolution of German: the case of Gothic*, in « Modern Language Notes », LXII (1947).