S. ZUBRZYCKI (Wrocław)

CONCERNING YULE'S CHARACTERISTIC OF STYLE

Statistical linguistics are concerned with numerical characteristics of style. Such characteristics are based on counts of occurences of particular words in a text. We aim to give here some comments on a characteristics of that kind, first introduced by Yule [5] and then considered by Herdan in [1] and [2], and described in his monograph [3]. We shall give an explanation, other than Herdan's, why these characteristics are reasonably constant for samples of different sizes. It will be shown, namely, that the constancy in question can be reduced to the known properties of samples drawn from an urn containing many kinds of balls. Occasionally we shall discuss a difference between Yule's characteristic and its modification proposed by Herdan, which was not explicitly stated in [2]. The difference consists in neglecting the terms of order $1/k$, where k is the number of different words in a text. This explains the visible although not very great divergence between the last two columns in a table in [1], reproduced also in [2] and [3], giving the two characteristics computed by Herdan for samples drawn from Pushkin's *The Captain's Daughter*.

Yule's characteristic can be defined as follows: Let be k — the number of different words in a given text, $p_1, \ldots, p_k$ — the frequencies of particular words in the whole text, *i. e.* the ratios of the number of occurences of a word by the number of all words in the text. Take from the text a sample counting n words. Let us denote by $a_i^{(n)}$ the number of occurences of the i-th word in our sample. If we put

$$S = \sum_{i=1}^{k} (a_i^{(n)})^2, \tag{1}$$

we can define Yule's characteristic K by the relation

$$K \cdot 10^{-4} = \frac{S}{n^2} - \frac{1}{n}. \tag{2}$$

When the manner of sampling is chosen, K becomes a random variable. We are now going to prove that if for each $i = 1, 2, \ldots, k$, the relation

(3) $a_i^{(n)}/n$ tends in probability to p_i for increasing n,

then [1]

(4) K tends in probability to $10^4 \sum_{i=1}^{k} p_i^2$ for increasing n.

This explains the remarkable stability of K for samples of not too small size.

To prove (4) it is sufficient to note that the term $1/n$ on the right side of (2) tends to 0 if n increases and thus (4) follows immediately from (3).

The relation (3), which is basic for all our comments, is a well-known consequence of the law of large numbers both for random sampling from a finite population (our text) with replacement and for random sampling without replacement. It can also be justified for samples composed of words following one another in the text, provided some conditions are fulfilled, *e. g.* that the text can be regarded as a realization of a stationary ergodic chain. The numbers p_i are then to be conceived as probabilities of particular words in that chain.

Herdan (cf. [1]) defines a somewhat different characteristic. Denoting by f_x the number of words in a sample which occur equally x times in it (in our previous notation f_x is the number of indices i such that $a_i^{(n)} = k$), he defines the mean, the standard deviation and the coefficient of variation of the number of occurrences of a word in a sample by the formulae

$$(5) \qquad M = \frac{\sum_{x=1}^{n} x f_x}{N^{(n)}}, \qquad s = \sqrt{\frac{\sum_{x=1}^{n} x^2 f_x}{N^{(n)}} - M^2}, \qquad v = s/M$$

where $N^{(n)}$ is the number of different words in a sample. Then he takes for a characteristic of style the expression

$$(6) \qquad H = \frac{v}{\sqrt{N^{(n)}}} = \frac{s/\sqrt{N^{(n)}}}{M}$$

and interpretes it as a ratio of the standard deviation of the sampling mean to the mean. Owing to the easily deducible equality

$$(7) \qquad \frac{v^2}{N^{(n)}} \frac{\sum_{x=1}^{n} x^2 f_x}{\left(\sum_{x=1}^{n} x f_x\right)^2} - \frac{1}{N^{(n)}},$$

[1] After having written this note I learned that the relation (4) was also found by Herdan (see [4], p. 71).

he explains why the characteristic H remains reasonably constant for samples of any size, arguing as follows: the second term on the right side of the formula (7) becomes negligibly small for great $N^{(n)}$, while the first term does not depend on n.

Here are some remarks about that reasoning. First of all let us note that the first term on the right side of (6) does not depend indeed explicitly upon the number $N^{(n)}$ of different words in a sample, but it depends on n, because the coefficients f_x do. Then the second term on the right side of (7) need not be negligibly small as compared with the first one. This can be seen as follows. First, the second term in question tends to $1/k$ and not to zero, except when the vocabulary is infinite. Secondly, we have the relation

$$(8) \qquad H^2 \text{ tends in probability to } \sum_{i=1}^{k} p_i^2 - \frac{1}{k}, \text{ if } n \text{ increases,}$$

which explains the stability of H. Thirdly, the limit in (8) becomes zero if $p_i = 1/k$, $i = 1, 2, \ldots, k$. Therefore, if we see in practical computations that the two characteristics do not differ very much we must regard it as a property of a text.

To prove (8) we have only to take into account the relations (3) d the following chain of equalities:

$$\frac{\sum_{x=1}^{n} x^2 f_x}{\left(\sum_{x=1}^{n} x f_x\right)^2} = \frac{\sum_{x=1}^{n} x^2 f_x}{n^2} = \frac{\sum_{i=1}^{k} \left(a_i^{(n)}\right)^2}{n^2} = \sum_{i=1}^{k} \left(\frac{a_i^{(n)}}{n}\right)^2.$$

References

[1] G. Herdan, *A new derivation and interpretation of Yule's characteristic K* Zeitschrift für angewandte Mathematik und Physik 6 (1955), No 4, p. 332-334.

[2] — *Ein neuer statistischer Parameter*, Zeitschrift für angewandte Mathematik und Mechanik 36 (1956), No 1/2, p. 72.

[3] — *Language as choice and chance*, Groningen 1956.

[4] — *An inequality between Yule's characteristic K and Shannons entropy H*, Journal of Applied Mathematics and Physics 9 (1958), p. 69-73.

[5] G. U. Yule, *Statistical study of vocabulary*, Cambridge 1944 (quoted after [1]).

MATHEMATICAL INSTITUTE OF THE POLISH ACADEMY OF SCIENCES

Received May, 19. 1958

S. ZUBRZYCKI (Wrocław)

O YULE'A CHARAKTERYSTYCE STYLU

STRESZCZENIE

W lingwistyce statystycznej rozważa się liczbowe charakterystyki stylu. Charakterystyki takie oparte są na obliczeniach liczby wystąpień poszczególnych słów w badanym tekście. Nota poświęcona jest dyskusji nad dwiema takimi charakterystykami. Jedna z nich pochodzi od Yule'a [5], a jej modyfikacja od Herdana [1]. W nocie podano probabilistyczne wyjaśnienie, inne niż u Herdana [1]-[3], tego, że wartości liczbowe tych charakterystyk praktycznie nie zależą od wielkości próbki. Podkreślono również różnicę między omawianymi charakterystykami.

С. ЗУБЖИЦКИЙ (Вроцлав)

О ХАРАКТЕРИСТИКЕ СТИЛЯ ПО ЮЛЮ

РЕЗЮМЕ

В статистической лингвистике рассматриваются численные характеристики стиля. Они основываются на подсчете числа выступлений отдельных слов в исследуемом тексте. Настоящая заметка посвящена дискуссии над двумя такими характеристиками. Одна из них представлена Юлем [5], а ее видоизменение Гердаком [1]. В заметке приводится вероятностное выяснение отличное от Гердэна [1]-[3] того, что численные величины этих характеристик не зависят от величины выборки. Подчеркнута также разница между оговоренными характеристиками.
