© NADEZHDA N. ZHURAVLEVA
natnicbel@gmail.com

УДК 81

APPLICATION OF QUANTITATIVE METHODS TO THE ANALYSIS OF AUTHOR'S STYLE AND THE SOLUTION OF THE PROBLEMS OF ATTRIBUTION

SUMMARY. The article presents a brief overview of the quantitative methods of linguistic analysis of style in the diachronic aspect. It considers the works of notable author's that deal with the application of statistical and computer methods to the solution of the problems of text attribution.

KEY WORDS. Style, attribution, quantitative methods, statistical methods, corpus of texts, concordance

The main objective of the present article is to make a brief overview of the development of quantitative methods in linguistics and find out how researchers have applied them to find peculiarities of an author's style and to solve problems of attribution.

Before the analysis of the development of quantitative methods it is necessary to consider the definition of style.

According to the definition of G. Herdan, style is the general characteristics of a person's way of expressing himself in language. "Style" is understood as a subconscious factor which the writer cannot but obey, and implies that linguistic expression is less a deliberate choice of words than it would appear at first sight [1; 12]. A person is unconscious of his style, and it can be identified as surely as fingerprints, provided that he does not deliberately set out to disguise it [2; 54].

We will take one more, simplified definition of style suggested by Werner Winter. A style may be said to be characterised by a pattern of recurrent selections from the inventory or optional features of a language. Various types of selections can be found: complete exclusion of an optional element, obligatory exclusion of a feature optional elsewhere, varying degrees of inclusion of a specific variant without complete elimination of competing features [3; 3].

Thus, from these two definitions of style we conclude the following: style is a pattern of recurrent selections characterising a person's way of expressing himself in language. This definition justifies the application of quantitative methods to the analysis of style.

There are various quantitative methods of author's stylistic analysis. We are going to consider them in a diachronic aspect.

The history of modern statistical stylistics apparently begins in the middle of the 19th century when the English mathematician Augustus de Morgan in 1851 made a suggestion that the styles of different authors could be distinguished by means of hidden statistical characters. His suggestions were made about some problems of Greek prose: he argued that the average length of words by an author might well prove to be a characteristic trait of a writer's style. However, as we know, de Morgan himself didn't make any actual counts [quat. from 4; 368].

There was also a group of mid-nineteenth century scholars developing a technique called "stylometrics". They counted the numbers of repetitions of certain words and the variations of metre in verse. The researchers presented their results as average or percentages. The school reached its height with the foundation, about 1874, of the New Shakespearean Society. Among the members of the Society were F.G. Freary ("On metrical tests applied to dramatics poetry I. Shakespeare ", 1874), J.K. Ingram (1874), F.W. Furnival (1887). The main result of their work was the discovery of a slow but steady change in style during the twenty-two years over which Shakespeare wrote thirty-six plays starting in 1589 when he was 26 years of age an ending in 1612 when he was 48 [quat. from 4:369].

The term "stylometry" (stylometrics) was also used by a German researcher W. Dittenberger (1880) who made an attempt to solve the task of attribution and chronology of the dialogues of Plato. He examined the frequencies of words, mainly function words, in Plato's texts. Later on, his investigations on various material were continued by E. Zeller (1903), F. Cada (1901) and C. Ritter, the latter comparing Plato and Goethe statistically [quat. from 4:369].

The development of ideas of de Morgan was the work of T.C. Mendenhall, an American geophysician, who, between 1887 and 1901, studied word length in English. He realized that the distribution of words of different lengths gave more opportunities for comparison of styles than the simple arithmetic mean proposed by de Morgan. Mendenhall's first paper entitled "The Characteristic Curve of Composition" (1887) was an outstanding advance towards the present-day stylostatistical approach. He investigated the difference between the literary styles of Dickens and Thackerey insofar as the word-length distribution was concerned and gave also examples of other writings in modern and classical languages. All his results were shown in the form of graphs ("word spectrums" as he called them), unfortunately without the lists of original numbers. In a later article (1901) Mendenhall used word-length frequency distributions in a study of the authorship of Shakespeare's plays. He showed that in every single count from Shakespeare's plays were more words with four letters than three. In comparison, Bacon had more three-letter words than four-letter words. Bacon had also a distinctly higher proportion of longer words than Shakespeare.

Mendenhall's published works seemed to have attracted little attention at that time. In this early period of statistical stylistics only few investigations with the help of statistical methods can be reported, e.g. the study by L.A. Sherman (1888) on sentence-length in English prose. H.A. Parker (1896) studied sentence legth in two works be Carlyle. C. Hildreth (1897) made a new contribution to the Shakespare-Bacon controversy. W. Lutostawski (1897) used statistical methods in establishing the chronology of Plato's dialogues, L. Fank

(1909) wrote on frequency of colour terms in Goethe's works. P. Partzinger (1911) studied the evolution of Ciciero's style [quat. from 4; 370].

In Russia N.A. Morozov raised the problem of differentiation between plagiarisms and the original works of notable authors. In 1915 he published the article "Linguistic spectrums". The authors before him based mainly on the frequency of the main parts of speech. Applying simple quantitative methods N.A. Morozov considered the frequency of functional parts of speech and their variations in individual texts [5].

In the twenties of the 20th century only a few serious stylostatistic works can be reported, e.g. by R.E. Parker (1925), Z.E. Chandler (1928), M. Parry (1928), and particularly, A. Busemann (1925), the inventor of the so-called Verb-Adjective-Ratio [quat. from 4:371].

In the thirties a new step was made in the use of statistical methods in stylistics by K.V. Fletcher (1934) who examined the evolution of the style of Spencer, G.V. Bolling (1937) with a critical essay on the statistical investigation of Homer's language, and finally classics of statistical stylistics such as J.B. Carroll (1938) who raised the problem of "diversity" in a vocabulary, and U.G. Yule (1938) with his first study on the sentence-length distribution as a statistical characteristic of style .

With U.G. Yule the real start of the application of modern statistical methods in stylistics begins. Since then the study of stylistic problems with the help of statistical methods has spread throughout the world. There was a rapid growth of interest in statistical linguistics, especially in the period of the sixties and the seventies (J.B. Carroll, G. Herdan, H.H. Somers, Ch. Muller, B. Kelemen, L.T. Milic, J. Mistrik, L. Doležel, C.B. Williams, B.N. Golovin, J. Kraus, M.N. Kozhina, etc.) [quat. from 4:371].

During that period various ideas of the analysis of author's style appeared.

According to Lubomir Doležel, the foundations of the statistical theory of style can be summarized in a simple statement: style is a probabilistic concept. Doležel thinks that this or that characteristic cannot be unequivocally determined by an author's habits. The probability distribution of a characteristic is rather a tendency. In this way, a probabilistic conception enables us to describe style not as a fixed habit, but rather a preference for one or another mode of expression. The overall character of style is called forth by the *degree* of presence (or absence) of a certain mode of expression, rather than by its exclusive use (or complete suppression). Thus, a probabilistic approach reveals the flexible character of stylistic features; these features resist any description in terms of necessity or in terms of strict rules and prohibitions [6; 11]. For this purpose it is necessary not only to compare several works of one author but, as suggested by H.H. Somers, but also analyze the differences between the works of known authors and the text under test [7; 128].

With the introduction of computers in the 1960s it became possible to use them in linguistic research, particularly, in the solution of the problem of authorship and author's style. Attracting linguists' attention to a computer is linked with its ability to store large amounts of information, in our case, corpus, and to find uses of words, word groups, repeated syllables, and so on. So, a computer can process large amounts of information in a fraction of a second. S.Y. Sedelow and W.A. Sedelow

introduced the new term "computational stylistics" which is understood as a quantitatively rigorous and intense study of pattern of style in natural language. Computational stylistics has immediate, practical implication for work in areas ranging from machine translation and automatic abreacting to social sciences and humanities. One of its uses is undoubtedly investigation of author's style and problems of attribution [8; 1].

M.H.T. Alford in his article "Computer assistance in language learning and in authorship identification" tries to solve the problem of attribution in the following way: he comes to the conclusion that words which appear as low frequency in general count are often high frequency in a local count. Thus, once a general low-frequency word has occurred in a text, its immediate future text-coverage is likely to be more than ten times that predicted by the general count [9; 84-85]. This conclusion enables us to state that one can find the stylistic peculiarities of an author on the basis of word frequency

Similar conclusions were made by G. Herdan in «Quantitative Linguistics»: style can be characterized by a constant ratio between uniformity and diversity of word frequencies[10; 71]. Herdan spoke about the following ratios: special vocabulary/ total vocabulary, special occurrences/ total occurrences, special vocabulary/total occurrences [1; 20-22].

With the beginning of the use of computers in linguistic research the problem of constructing concordances appeared. Concordance is a list of contexts in which we have a particular word or sequence of words. Concordance program is a basic tool of corpus linguistics that converts electronic text into a database that can be studied [11]. With the help of these programs we can look for words, phrases, parts of words, this program is able to create a list of collocations, as well as collect data on frequency of use [12; 57].

One of the scientists dealing with the problem of concordances was D. Ross. He explained the problem of constructing concordances by the fact that it is hard for a computer to determine parts of speech. He criticized Ellegard's frame approach that suggested determining them by the word order. Thus, words which fell between a determiner and another function word were called nouns, and the adjective category was defined as those words framed by a determiner and a noun. The difficulty arises when a word has multiple categories. Instead Ross suggested his own concordance program "EYEBALL" which marks only the words under focus and each new phrase or clause is treated in isolation from the others. Another important difference between EYEBALL and other computing procedures for stylistic analysis is the inclusion of the functional labels, which makes the syntactic description considerably more complete than is possible with only categorical labels [13; 88].

In the 70-90s, more and more researchers became interested in the use of computer data processing in the analysis of texts, in syntactical as well as in grammatical, lexical aspects.

Mandatory application of automatic data processing is in the basis of the works by of U.V. Sidorov, I.O. Tarnopolskaya, D.V. Khmelev. In the study of texts, organized under the leadership of L.V. Milov, attribution of texts is done by constructing graphs of "strong ties" according to the matrix of the frequency of the pair occurrence of grammatical classes of words with the help of a special computer program [14].

One of the most modern Rissian linguists involved in the statistics is G.Y. Martynenko. In 1988 he wrote the monograph "Foundations of stylometrics" and for more than twenty years he has been practising statistical methods in linguistics. The most recent of his research is the study of the theory of so-called "golden ratio", come up with the Pythagoreans, in linguistics. Considering the syntactic structure in terms of measures of syntactic complexity, rhythmic structures, the ratio of single and multiple tokens, he concludes that all of are regulated by the law of "golden ratio". Here it may seem remarkable that, with the help of the law he tried to consider more than one level (phonemic, morphological, syntactic) which allows to analyze style from different angles [15].

Another modern Russian linguist engaged in statistical methods of the attribution of the text is M.A. Marusenko. He created the idea of image recognition. He divided the procedure of the attribution into the three relatively independent phases:

- The formation of literary-critical attribution hypothesis,
- The verification of literary-critical attribution hypothesis using the theory of image recognition,
- Interpretation of test results of the attribution hypothesis.

In this paper, statistical and probabilistic methods of analysis of language and style are used by the author for testing the attribution hypothesis [16; 25].

Thus, we understand style as a probabilistic concept, as a pattern of recurrent selections characterising a person's way of expressing himself in language, mainly unconscious. For the analysis of style we cannot look for exclusive use or complete suppression of a characteristic it is necessary to determine its *degree* of presence or absence. Quantitative methods and concordance program will be more convenient for that purpose.

REFERENCES

1. Herdan, G. The advanced theory of language as choice and chance. Berlin: Springer-Verlag, 1966. 365 p.

2. Booth, Andrew D.; Brandwood, L.; Cleave, J.P. Mechanical resolution of linguistic problems. London: Butterworks scientific publications, 1958. 306 p.

3. Winter, Werner. Styles as dialects // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company, INC, 1969. P. 3-9.

4. Tuldava, Juhan. Stylistics, author identification // Quantitative linguistics: an international handbook / Edited by Reinhard Köhler, Gabriel Altmann, Raĭmond Genrikhovich Piotrovskiĭ. Berlin, New York: de Gruyter, 2005. P. 368-387.

5. Morozov, N.A. A new instrument for objective investigation of the documents. (Linguistic spectra as the means to distinguish the true from the plagiarized works of a famous author and to determine their age). URL: http://www.textology.ru/libr/Morozov.htm

6. Doležel, Lubomir. A framework for the statistical analysis of style // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company, INC, 1969. P. 10-25

7. Somers H.H. Statistical methods in literary analysis // The computer in literary style. Introductory essays and studies / Edited by Jacob Leed. Kent, Ohio, USA: Kent State University press, 1966. P. 128-140.

8. Sedelow, Sally Yeates; Sedelow, Walter A., Jr. A preface to computational linguistics // The computer in literary style. Introductory essays and studies / Edited by Jacob Leed. Kent, Ohio, USA: Kent State University press, 1966. P. 1-13.

9. Alford M.H.T. Computer assistance in language learning and in authorship identification // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company, INC, 1969. P. 77-86.

10. Herdan G. Quantitative linguistics. London: Butterworths, 1964. 284 p.

11. McEnery T., Wilson A. Corpus linguistics [Электронный ресурс]. URL: http:// bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus1/1fra1.htm.

12. Philipenko, T.V. Using the methods of corpus linguistics in the analysis of the semantics of idioms // Vestnik MGU. Ser.19. Linguistics and Intercultural Communication. 2004. № 1. P. 84-88.

13. Ross D. Beyond the concordance: algorithms for description of English clauses and phrases // The computer and literary style / Edited by A. J. Aitken, R.W. Bailey, N. Hamilton- Smith. Edinburgh: University Press, 1973. P. 85-99.

14. From Nestor to Fonvizin. New methods for the determination of authorship. / Edited by L.V. Milov. M.: «Progress», 1994. 378 p.

15. Мартыненко Г.Я. Золотое сечение в нумерологии текста. URL: http:// numbernautics.ru

16. Marusenko, M.A. The attribution of anonymous and pseudonymous literature by pattern recognition methods. L.: Leningrad State University Publishing House, 1990. 168 p.