BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **16**, No 3
Sofia • 2016 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2016-0043

The Distribution of Semantic Fields in Author's Texts

Bohdan Pavlyshenko

Ivan Franko, Lviv National University, Ukraine
E-mail: b.pavlyshenko@gmail.com

Abstract: *The paper describes the analysis of frequency distribution of semantic fields of nouns and verbs in the texts of English fiction. To such distributions, we applied Shapiro-Wilk test. The null hypothesis of normal distribution of semantic fields frequencies in the array of texts under analysis is rejected for some semantic fields. This makes it possible to consider the frequency distribution of semantic fields as a categorized mixture of normal distributions. As a factor of categorization, we chose text authorship. We divided the author's categories with rejected hypothesis of normal distribution into subcategories with normal distribution. Paired Student's t-test for the distributions of semantic fields in the texts of different authors revealed a measure of authorship representation in the structure of semantic fields. The analysis of the results showed that the author's idiolect is represented in the vector space of semantic fields. Such a space can be used in the analysis of the authorship and author's idiolect of texts.*

Keywords: *Frequency distribution, semantic fields, Shapiro-Wilk test.*

1. Introduction

In analysis of text arrays a vector model of text documents is used, according to which the documents are considered as vectors in some vector space, formed by quantitative characteristics of words (P a n t e l and T u r n e y [8]). As a quantitative characteristics, the frequencies of keywords are widely used. One of the problems of such an approach is a large dimension of text documents space, which is caused by the size of the vocabulary of text array under analysis. A promising approach to solve this problem is the use of vector space with a basis formed by quantitative

characteristics of word associations, in particular semantic fields. A semantic field is a set of words that are united under some common concept. The examples of semantic fields can be the field of motion, the field of communication, the field of perception, etc. The number of semantic fields is significantly smaller than the size of a word dictionary, and it reduces the amount of necessary calculations. Similar objects are the semantic networks that describe the relationships among different concepts. An example of a lexicographic computer system, which represents the semantic network of links between words, is a WordNet system developed at Princeton University (Fellbaum [2]). This system is based on an expert lexicographic analysis of semantic structural relationships that describe the denotative and connotative characteristics of dictionary word composition. The paper (Gliozzo and Strapparava [3]) considered the concept of semantic domain, which describes certain semantic areas of various issues discussed, such as economics, politics, physics, programming, etc. The algorithms of clusterization and classification are often used in data mining (Sebastiani [13]; Manning, Raghavan and Schütze [6]). In Pavlyshenko [9], the use of Naive Bayesian classifier (NB) and the classifier by the *k* Nearest Neighbors (kNN) in classification semantic analysis of author's texts of English fiction has been analyzed. The author's works are considered in the vector space the basis of which is formed by the frequency characteristics of semantic fields of nouns and verbs. Highly precise classification of author's texts in the vector space of semantic fields indicates the presence of particular spheres of author's idiolect in this space which characterizes the individual author's style. In Pavlyshenko [10], the analysis of possible differentiation of the author's idiolect in the space of semantic fields has been described. The analysis showed that using the vector space model with the basis of semantic fields is effective in the cluster analysis algorithms of author's texts in English fiction. The study of the distribution of author's texts in the cluster structure showed the presence of the areas of semantic space that represent the idiolects of individual authors. Such areas are described by the clusters where only one author dominates. The clusters, where the texts of several authors dominate, can be considered as areas of semantic the similarity of author's styles.

In this paper, we study the frequency distributions of the semantic fields of nouns and verbs in the texts of English fiction. We consider such distributions as categorized mixtures of normal distributions. The main aim of this work is to study the frequency distribution of the semantic fields of nouns and verbs in the texts of English fiction as an additional factor for the investigation of author's style. In Section 2, we consider the theoretical model of text documents in the space of semantic fields, probability distribution of the author's style in the documents of text array. In Section 3, we show the results of our studies. In Section 4, we summarize our study and make conclusions.

2. The model of text documents in the space of semantic fields

Let us consider a model based on a set theory, which describes a set of text documents and semantic fields. We describe a set of text documents as

(1) $$D = \{\ d_j \mid j = 1,\ 2, ..., N_d\}.$$

We introduce a set of semantic fields

(2) $$S = \{\ s_k \mid k = 1,\ 2, ..., N_s\}.$$

Then we form a matrix of a feature-document type where the features are the frequencies of semantic fields in the documents:

(3) $$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,\, j=1}^{N_s, N_d}.$$

The frequencies of semantic fields p_{kj}^{sd} are defined as the sums of word text frequencies that are included into these semantic fields. The values of these frequencies are normalized so that their sum for each document is equal to 1. The vector

(4) $$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right)^{\mathrm{T}}$$

displays the document d_j in N_s-dimensional space of text documents. The introduction of the semantic fields space not only reduces the size of the problem of texts analysis, but also introduces a new basis for text descriptions. One of possible models explaining such a result could be a mixture of normal distributions model (Hofmann [5], Hansen at al. [4], Zhai, Velivelli and Yu [14], Rosen-Zvi at al. [12], Mei and Zhai [7], Benaglia at al. [1]). According to this model, the distribution of frequencies is considered as a sum of functions of normal distributions of semantic fields with coefficients. Each such function describes the frequency distribution of semantic fields in the documents of given category. As a category of documents, we consider the text authorship. Some distributions where the null hypothesis of the normal distribution was rejected can be similarly considered as a mixture of normal distributions for author's subcategories by given semantic field. Given the unique nature of semantic fields frequency distribution in the texts of various author's categories, one can construct a probable model of author's styles distribution in the documents of text array. In this model, semantic fields can play a role of hidden parameters. Such a model can be represented as a probability distribution of the author's style in the documents of text array.

(5) $$P(\mathrm{Style}_j^a, d_i) = P(d_i) \cdot P(\mathrm{Style}_j^a \mid d_i),$$

where

$$P(\mathrm{Style}_j^a \mid d_i) = \sum_k^{N_s} f_k^s(\mathrm{Style}_j^a \mid p_k^s) \cdot f_k^s(p_k^s \mid d_i),$$

$f_k^s(p_k^s \mid d_i)$ is the frequency of semantic fields in the analyzed document d_i, $f_k^s(p_k^s \mid d_i) = p_{ki}^{sd}$. The value $f_k^s(\mathrm{Style}_j^a \mid p_k^s)$ can be found on the basis of constructed functions of the semantic field frequency distribution in the documents

of given category. The semantic fields, as hidden parameters, play a role of style-dividing factors in classification analysis.

3. Experimental part

For the calculations, we used R software environment (R Core Team [11]). For the experimental study of text documents clustering in the space of semantic fields, we chose a text base containing 503 literary works of 17 authors (A. K. Doyle (1), A. Trollope (2), Ch. Dickens (3), E. Gaskell (4), E. Lytton (5), G. Meredith (6), H. Wells (7), J. Conrad (8), J. Galsworthy (9), J. London (10), M. Twain (11), R. Kipling (12), R. Stevenson (13), T. Hardy (14), W. Colllins (15), W. Scott (16), W. Thackeray (17)). For the semantic space generation we chose the words grouped by the semantic fields of nouns and verbs in the semantic network WordNet (Version 2.1) (Fellbaum [2]). The semantic fields in the WordNet network (**http://wordnet.princeton.edu**) are represented as lexicographic files. In our studies we have used the semantic fields of nouns and verbs. The semantic fields of nouns consist of 26 lexicographic files with selected 54464 words. The semantic fields of verbs contain 15 lexicographic files with selected 9097 words. The derivative forms of words were also included into the semantic fields. Lexicographic files WordNet for nouns and verbs have the names that define the semantic core of these fields: noun.tops(1), noun.act(2), oun.animal(3), noun.artifact(4), noun.attribute(5), noun.body(6), noun.cognition(7), noun.communication(8), noun.event(9), noun.feeling(10), noun.food(11), noun.group(12), noun.location(13), noun.motive(14), noun.object(15), noun.person(16), noun.phenomenon(17), noun.plant(18), noun.possession(19), noun.process(20), noun.quantity(21), noun.relation(22), noun.shape(23), noun.state(24), noun.substance(25), noun.time(26), verb.body(27), verb.change(28), verb.cognition(29), verb.communication(30), verb.competition(31), verb.consumption(32), verb.contact(33), verb.creation(34), verb.emotion(35), verb.motion(36), verb.perception(37), verb.possession(38), verb.social(39), verb.stative(40), verb.weather(41).

The examples of the distributions of the semantic fields frequencies, represented with the help of a boxplot type of graphics, are shown on Fig. 1. The box plot allows us to receive visual information about semantic fields distributions. The thick line in the box denotes median, the top and bottom box borders denote first and third quartiles, the horizontal lines denote the range of values of semantic fields frequencies, small circles denote outliers.

As the results presented show, the main features of frequency distributions can be significantly different for the collections of different authors. The examples of the semantic fields frequency distributions in the text arrays of some authors are shown on Fig. 2.

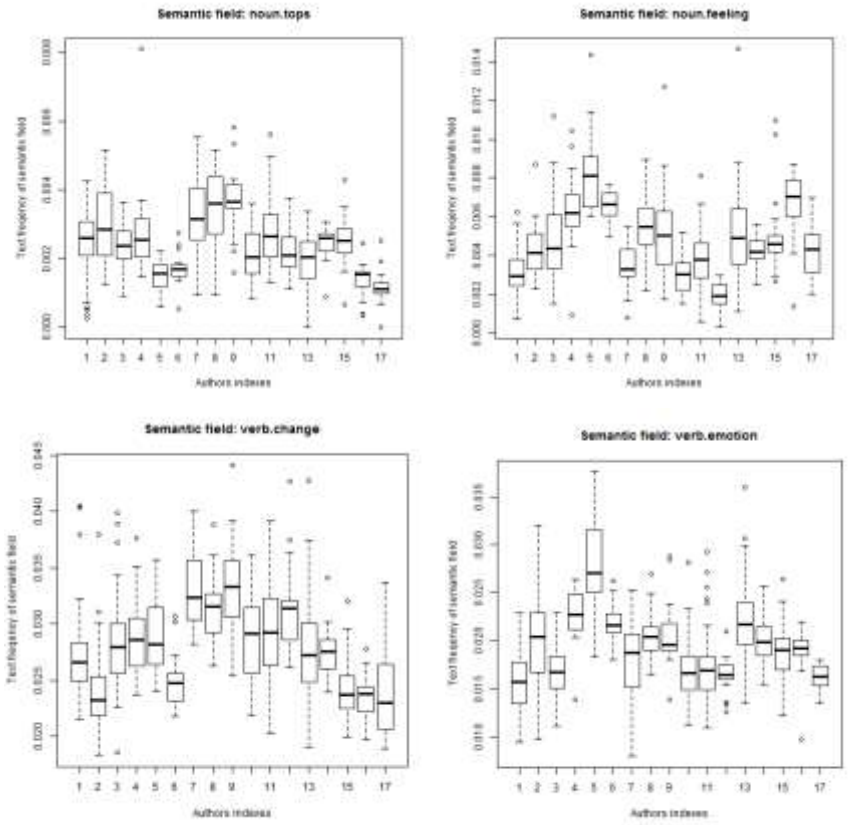

Fig. 1. The examples of distributions of semantic fields frequencies by authors

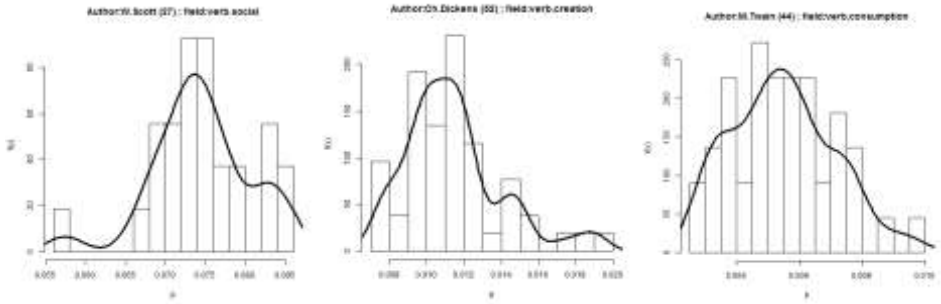

Fig. 2. The examples of semantic fields distributions in the text arrays of some authors (the number of considered texts is shown in brackets after the author's name)

To detect the semantic fields with the style-dividing potential, we calculate the standard deviation for semantic fields frequencies averaged by author's categories. The results of obtained calculations are shown on Fig. 3.

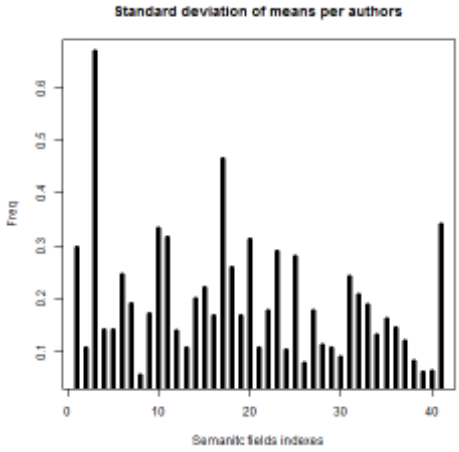

Fig. 3. The standard deviation of semantic fields means

Let us consider the frequency distributions of semantic fields in the analyzed array of text documents. For checking up the null hypothesis of normal distribution, we use the standard Shapiro-Wilk test with the significance level of 0.05. We conduct the test in R software environment. On the basis of the results of the conducted test, one can reject the null hypothesis for almost all frequency distributions of semantic fields. Our next step is to carry out the Shapiro-Wilk test for frequency distributions of semantic fields for each author's category of texts. For the authors under study we received the following values of numbers of semantic fields with non-normal distributions: A. K. Doyle – 21, A. Trollope – 26, Ch. Dickens – 30, E. Gaskell – 15, E. Lytton – 19, G. Meredith – 15, H. Wells – 13, J. Conrad – 12, J. Galsworthy – 15, Jack London – 23, Mark Twain – 19, R. Kipling – 14, R. Stevenson – 23, T. Hardy – 16, W. Colllins – 24, W. Scott – 17, W. Thackeray – 18. The distributions where the null hypothesis about the normal distribution was rejected can be considered as a mixture of normal distributions for author's subcategories by given semantic field. To calculate the parameters of the distributions, we use the realization of the EM algorithm of "mixtools" package for the R environment. Let us consider the distribution of the semantic fields in the set of texts of one author. The non-normal frequency distribution of semantic fields can be represented as mixture of normal distributions. Fig. 4 shows the calculated example of the histogram and a mixture of normal distribution of the semantic field noun.animal for A. Doyle's texts. The mixture model explains the existence of text subgroups in the observed set of author's texts. These subgroups are defined by the distribution of the semantic fields.

So, non-Gaussian distributions of the semantic field frequencies can be described on the basis of the mixture model of categorized distributions of the semantic field frequencies. Since we chose the existing classification of texts by authors as the categories, in some cases the distribution of semantic fields frequencies in the categories may be non-Gaussian. In a case like that, the author's category can be divided into extra subcategories with Gaussian distribution.

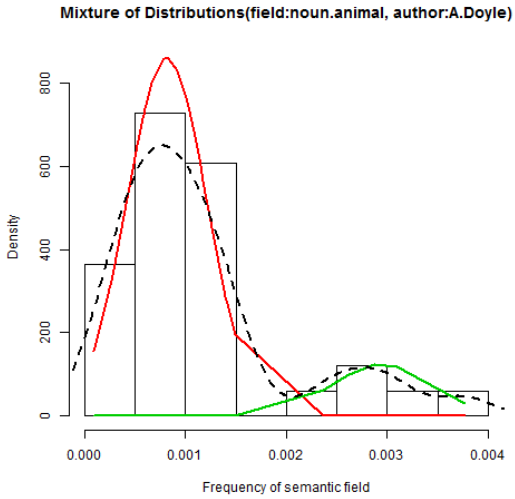

Fig. 4. Histogram and mixture of normal distribution of chosen semantic fields in the set of texts of one author

Let us assume that the semantic fields frequencies reflect the author's idiolect in the text author's categories under analysis. It can be detected by comparing the frequency distributions of some semantic field in the texts of various author's categories. If these distributions are different for different authors, so they reflect the author's idiolect. To compare the changes of frequencies of different semantic fields, we will calculate the relative change of frequencies for each author's texts collection. Fig. 5 shows the changes of relative frequencies averaged by authors. Individual set of frequency changes for the texts of individual authors makes it possible to consider the vector space of semantic fields as low-dimensional space for classification algorithms and text arrays clusterization in the tasks of the analysis of author's idiolect.

Let us compare the means of two frequency distributions of the semantic fields of the texts of two authors. To do that we calculated the *p*-value of Student test. If *p*-value < 0.05, then the mean values of two investigated distributions are different, otherwise the hypothesis is accepted that such mean values are equal. As a result of applying Student test to all the pairs of author's texts sets, we receive N_{fields} matrixes with the dimensions $N_{\text{auth}} \times N_{\text{auth}}$ with *p*-values. Each matrix was calculated for each semantic field.

Each *p*-value in the matrix denotes the result of Student test between the distribution of chosen *k* semantic field in the texts set of *i* and *j* authors. Let us transform the received values to simplify the results. If *p*-value > 0.05, then we replace it by 0, otherwise we replace it by 1. Value 1 means that distributions under investigation are different, value 0 means that these distributions are the same. Then we calculate average value for each matrix. These average values describe the percent of distributions of semantic fields which are statistically different. If some quantitative characteristics has statistically different distributions in the texts sets of different authors, it means that that it can be considered as author's style defining

characteristic. For practical implementation of Student's t-test, we used the t.test() function from R package. Fig. 6 shows such author's style defining characteristic for some semantic fields. Obtained results showed that some semantic fields have high defining potential for differentiating author's style.

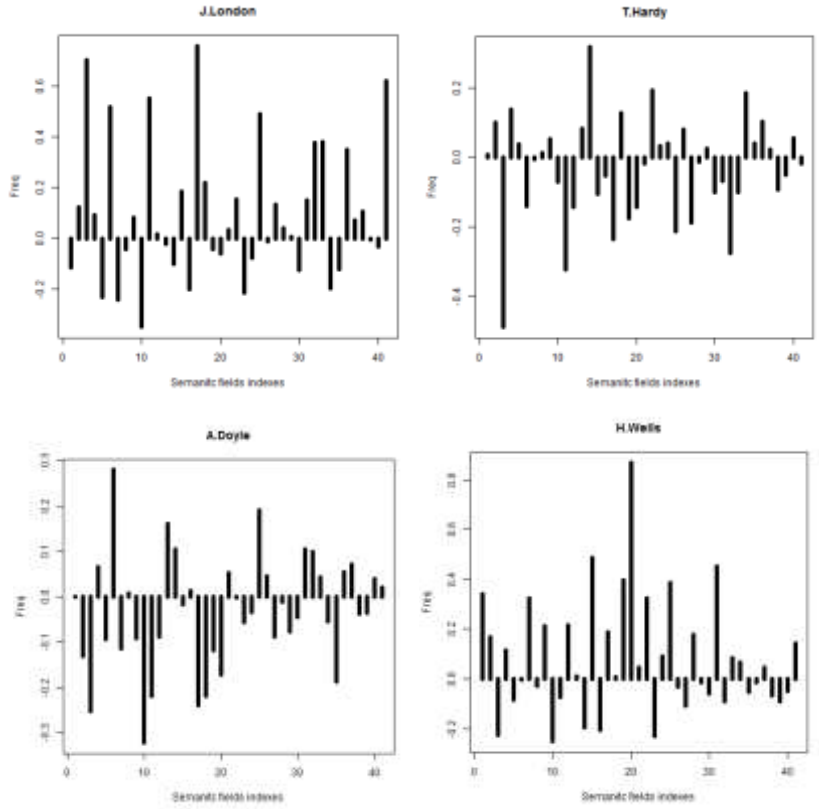

Fig. 5. Relative change of semantic fields frequencies in the sets of authors' texts

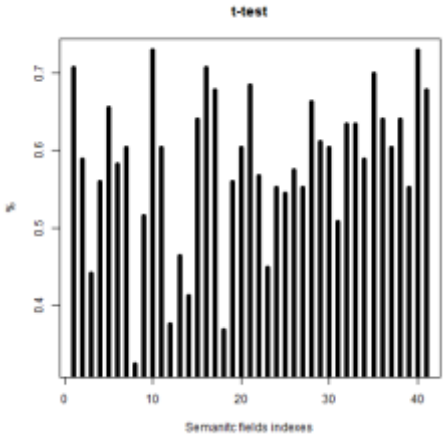

Fig. 6. Author's style dividing characteristics for semantic fields

5. Conclusion

In this paper, we investigated the frequency distribution of semantic fields of nouns and verbs in the texts of English fiction. The null hypothesis of normal distribution of semantic fields frequencies in the array of texts under Shapiro-Wilk test analysis is rejected for some semantic fields. This makes it possible to consider the frequency distribution of such semantic fields as a categorized mixture of normal distributions. As a factor of categorization, we chose text authorship. We divided the author's categories with rejected hypothesis of normal distribution into the subcategories with normal distribution. Paired Student's t-test for the distributions of semantic fields in the texts of different authors revealed the measure of authorship representation in the structure of semantic fields. The analysis of obtained results showed that the author's idiolect is represented in the vector space of semantic fields. Such a space can be used in the tasks of predictive analysis of the author's idiolect of texts. Some semantic fields have high dividing potential for differentiating of the author's style. As the results show, the distribtions of semantic fields can be considered as an additional factor for the structural investigation of author's texts.

References

1. Benaglia, T., D. Chauveau, D. R. Hunter, D. S. Young. Mixtools: An R Package for Analyzing Finite Mixture Models. – Journal of Statistical Software, Vol. **32**, 2009, No 6, pp. 1-29.
2. Fellbaum, C. WordNet. An Electronic Lexical Database. Cambridge, MA, MIT Press, 1998.
3. Gliozzo, A., C. Strapparava. Semantic Domains in Computational Linguistics. Springer, 2009.
4. Hansen, L. K., S. Sigurdsson, T. Kolenda, F. A. Nielsen, U. Kjems, J. Larsen. Modeling Text with Generalizable Gaussian Mixtures. – In: Proc. of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00), IEEE 2000, Vol. **6**, 2000, pp. 3494-3497.
5. Hofmann, T. Probabilistic Latent Semantic Indexing. – In: Proc. of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, August 1999, pp. 50-57.
6. Manning, C. D., P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
7. Mei, Q., C. Zhai. A Mixture Model for Contextual Text Mining. – In: Proc. of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, August 2006, pp. 649-655.
8. Pantel, P., P. D. Turney. From Frequency to Meaning: Vector Space Models of Semantics. – Journal of Artificial Intelligence Research, Vol. **37**, 2010, pp. 141-188.
9. Pavlyshenko, B. Classification Analysis of Authorship Fiction Texts in the Space of Semantic Fields. – Journal of Quantitative Linguistics, Vol. **20**, 2013, No 3, pp. 218-226.
10. Pavlyshenko, B. Clustering of Authors' Texts of English Fiction in the Vector Space of Semantic Fields. – Cybernetics and Information Technologies. Vol. **14**, 2014, Issue 3, pp. 25-36.
11. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.
https://www.R-project.org

12. R o s e n-Z v i, M., T. G r i f f i t h s, M. S t e y v e r s, P. S m y t h. The Author-Topic Model for Authors and Documents. – In: Proc. of 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, July 2004, pp. 487-494.
13. S e b a s t i a n i, F. Machine Learning in Automated Text Categorization. – ACM Computing Surveys, Vol. **34**, 2002, pp. 1-47.
14. Z h a i, C., A. V e l i v e l l i, B. Y u. A Cross-Collection Mixture Model for Comparative Text Mining. – In: Proc. of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, August 2004, pp. 743-748.