Formation of functional-role communication clusters based on morphological features of the verbal context

Vladimir Meshkov¹, Natalia Kochkovaya¹,*, and Irina Usova¹

¹Institute of Technologies (branch) of DSTU in Volgodonsk, 16, Mira avenue, Volgodonsk, 347386, Rostov region, Russia

Abstract. This work combines two approaches to the definition of the author's style: statistical and linguistic (method of morphological analysis). The average data of the authors obtained on the basis of statistical-morphological analysis are taken as an indications vector. The article shows that this technique allows to recognize the author's style of the work with sufficient accuracy, on an average about 88 percent. It is shown that the closest measure of proximity to this technique is the Euclid distance. The prospects for the development of this approach are outlined.

1 Introduction

Let us consider communication as a process of informational interaction of subjects and / or objects (for example, bots) in society. For interpersonal and, especially, for functional role interaction, the use of role varieties is characteristic. Moreover, the content and orientation of communication, the forms of its expression are determined by the partners role relations.

Recently the task of determining the type or role of the interlocutor and as a consequence the choice of a particular strategy for organizing a dialogue more and more often requires solving in the organization of communications (dialogue communication).

It becomes especially relevant when participating in communication functional-role interaction of artificial systems (bots).

In this paper we propose an approach to the formation of signs of the role of the verbal context based on the analysis of the vector of its morphological components. According to the authors point of view it will allow to solve the problems of automatic clustering of verbal context flows in the future and on this basis the choice of an appropriate dialogue scenario can be solved too.

The work presents the testing of the proposed approach based on the analysis of the Russian language literary works to define the author's style of the corresponding document. The choice of such an application area is due not only to the availability of appropriate tools (lemmatizer) but also the availability of a huge array of source texts.

Nowadays one of the urgent tasks in the field of text automatic classification is the
author's style determination. However, the question of the text analyses – what are the indications of the text – is not completely resolved.

There are different approaches to the definition of the author’s style since it is often difficult to formalize signs which are obvious to humans [1, 2, 3]. In addition, the perception of text is influenced not only by the author's style but also by the nuances of the readers’ subjective perception for example the emotional background evoked by a given work. The linguistic methods are the most effective because they are focused on solving such problems. However, the implementation of linguistic methods requires a very significant computational cost [4, 5, 6]. First of all this reason makes difficult to apply this approach in network technologies (search systems etc.), that is, where the problem must be solved on the wing.

In this paper an attempt is made to combine two approaches both statistical and linguistic (namely, the method of morphological analysis) to determining the author's style of an arbitrary work.

Morphological analysis [7] is one of the components of the linguistic processor and is used mainly to determine the grammatical meanings of words [8], such as parts of speech, gender, number, etc [9].

We intend to consider the quantitative composition and correlation of parts of speech in the text as the main characteristic of the author's style. The main idea of the study is the vector of morphological features of the text can be combined into a vector in a multidimensional space of values of morphological features. Then, such a vector (in other words a point in a given space) carries information about the author's style of the work. Moreover, it can be assumed that the vector also characterizes the literary work author. Therefore, having a database of vectors of morphological features of texts of various authors, it is possible to automate the definition of the author’s style (or the proximity of the style of an arbitrary work) to a certain famous writer.

Using the author's program of morphological analysis, Russian texts of various authors were processed. The program uses morphological modules from the ATP working group (Automatic Text Processing http://www.aot.ru) [10]. Since the lemmatizers presented by the ATP group in free access are focused on the processing of the Russian language [11] the works of authors of native Russian speakers were mainly considered.

E.g. the results of processing the works of three different authors (B Akunin “Turkish Gambit”, D Rubina “Babas Wind”, E Vodolazkin “Lavr”) are shown in the table 1. The table shows both the absolute and relative values of the morphological characteristics of the literary works.

Analyzing the values of relative values of morphological characteristics, we see that they differ among all authors to a great extent.

### Table 1. The values of the vectors of morphological characters of three authors.

|                      | B Akunin “Turkish Gambit” | D Rubina “Babas Wind” | E Vodolazkin “Lavr” |
|----------------------|----------------------------|-----------------------|---------------------|
| **Total amount of words in the text** | 11925                      | 15884                 | 21724               |
| **Number of masculine nouns** | 1433 0.1201677            | 1776 0.1118106        | 2082 0.0958387      |
| **Number of feminine nouns** | 856 0.0717819            | 1705 0.1073407        | 1995 0.0918339      |
| **Number of nouns of the neuter gender** | 455 0.0381551            | 614 0.0386552         | 1179 0.0542717      |
| **Number of masculine and feminine nouns** | 15 0.0012578            | 26 0.0016368          | 10 0.0004603        |
The differences are more clearly seen in the diagram below for the distribution of characters of these three works (figure 1). The differences are especially visible in adjectives, adverbs and imperatives.

If we consider a number of works by one author, we will see an obvious similarity and even coincidence of the values of the morphological parameters of the texts. For example table 2 shows the processing data of three works by contemporary author D. Sillov: “The Law of Pripyat”, “The Law of the Hunter”, “The Law of the Yakuza”.

![Figure 1](image-url)

**Table 2.** Values of vectors of morphological features of one author.

| Number of words in the text | «The Law of Pripyat» | «Law of the Hunter» | «The Law of the Yakuza» |
|-----------------------------|----------------------|---------------------|------------------------|
| 14708                       | 15166                | 15046               |
| Number of masculine nouns   | 1542                 | 1636                | 1683                   | 0.111857 |
| Number of feminine nouns    | 1104                 | 1191                | 1109                   | 0.073707 |
The table 2 very accurately reflects the almost coincidence of the coordinates of the vector of morphological characters for all three works of the author. The coincidences are visible on the diagram (figure 2). It can be seen that the graphics almost merge into one.

Fig. 2. The distribution diagram of one author works features.

Let’s consider how recognition of an author’s style is carried out in the considered methods.

An example of recognizing the author’s style for a fragment of the database formed on the basis of the work of seven authors (Strugatskys, Levitsky, Perumov, Pekhov, Akunin, Abramov, Prilepin) is shown in table 3.
Table 3. Fragment of a database of works with average values of vectors of morphological characters.

| Parameter                              | Strugatskys | Levitsky | Perumov | Pekhov | Akunin | Abramov | Prilepin | Levitsky new |
|----------------------------------------|-------------|----------|---------|--------|--------|---------|----------|--------------|
| Number of masculine nouns              | 0.11        | 0.15     | 0.13    | 0.14   | 0.14   | 0.12    | 0.11     | 0.16         |
| Number of feminine nouns               | 0.07        | 0.11     | 0.09    | 0.09   | 0.09   | 0.11    | 0.10     | 0.11         |
| Number of nouns of the neuter gender   | 0.04        | 0.04     | 0.03    | 0.05   | 0.04   | 0.04    | 0.03     | 0.04         |
| Number of masculine and feminine nouns | 0.00        | 0.00     | 0.00    | 0.00   | 0.00   | 0.00    | 0.00     | 0.00         |
| Number of verbs                        | 0.15        | 0.18     | 0.15    | 0.18   | 0.15   | 0.16    | 0.17     | 0.19         |
| Number of participles I                | 0.01        | 0.02     | 0.02    | 0.02   | 0.01   | 0.01    | 0.03     | 0.02         |
| Number of participles II               | 0.01        | 0.02     | 0.02    | 0.02   | 0.02   | 0.01    | 0.02     | 0.02         |
| Number of adjectives                   | 0.09        | 0.10     | 0.11    | 0.11   | 0.11   | 0.09    | 0.11     | 0.10         |
| Number of adverbs                      | 0.13        | 0.11     | 0.12    | 0.12   | 0.11   | 0.12    | 0.13     | 0.12         |
| Number of imperatives                  | 0.01        | 0.01     | 0.02    | 0.01   | 0.01   | 0.02    | 0.01     | 0.01         |
| Number of interjections                 | 0.06        | 0.05     | 0.05    | 0.05   | 0.06   | 0.06    | 0.05     | 0.06         |
| Number of prepositions                 | 0.09        | 0.12     | 0.08    | 0.10   | 0.10   | 0.12    | 0.10     | 0.12         |
| Number of conjunctions                 | 0.12        | 0.09     | 0.12    | 0.10   | 0.11   | 0.13    | 0.10     | 0.10         |
| Number of pronouns                     | 0.20        | 0.12     | 0.16    | 0.17   | 0.13   | 0.15    | 0.14     | 0.12         |
| Distance                               | **0.11**    | **0.011**| **0.078**| **0.059**| **0.061**| **0.070**| **0.061**| **0.011**    |
| Coincide                               | No          | Yes      | No      | No     | No     | No      | No       | No            |

There is a vector of morphological features of the determined author (the analyzed work) in the right column. The bottom line shows the results of the analysis. It can be seen that in this example the recognition is 100% (the author's style is Levitsky). In this example the distance between the ends of the vectors of morphological features of the text was determined based on the Euclid distance.

Similar studies were conducted for other proximity measures - “Chebyshev distance” and “city block distance – Manhattan” [12, 13]. Good results were obtained with the “city block distance” and practically acceptable with the “Chebyshev distance” [13].

Based on the research the following conclusions can be done: best of all the methods function for the works of writers of the late 19th and the early 20th centuries (“Russian classics”).

We also note that the most acceptable measure was the proximity “Euclid distance”.
On an average, this methods has shown a good practical result in recognizing the author's style (about 85-88%). It can also be noted that there is an error in the work of the author’s program related to the incorrect definition of the normal form of a word, on the basis of which a formal attribute is also determined. This is due to the inability to correctly determine the form of a word without parsing a sentence [14, 15]. However such a parsing requires additional significant computational resources it negates the advantages of the proposed methods. Incidentally the methods in our opinion is practically accurate, promising and efficient.

In the future the database should be expanded by including new authors and their works. It is also possible to conduct a cluster analysis of the obtained data to identify certain communities of authors following similar author’s styles.

References

1. E.I. Bolshakova, A.S. Sapin, *Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue”* (Russian State University for the Humanitie, Moscow, 2019) https://publications.hse.ru/en/chapters/315497919
2. S. Srinidhi, *Towards data science* (2019) https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058
3. L.K. Nelson, *Dlab Berkeley Edu* (2013) https://dlab.berkeley.edu/blog/ceos-and-automated-text-analysis
4. L. Jiang, E. Bolshakova, *Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue”* (Russian State University for the Humanitie, Moscow, 2019) https://easychair.org/publications/paper/4jj4
5. *Surceforge* (2020) https://sourceforge.net/projects/sem/
6. V. Vasilevski, L. Seliar-Cabral, M. Jose Araujo, *Automatic Analysis of Portuguese Verb Morphology Conference Paper* (2012) https://www.researchgate.net/publication/262327368_Automatic_Analysis_of_Portuguese_Verb_Morphology
7. 7.*Morphology (linguistics), Wikipedia* (2020) https://en.wikipedia.org/wiki/Morphology_(linguistics)
8. Semantic analysis (linguistics), Wikipedia (2020) https://en.wikipedia.org/wiki/Semantic_analysis_(linguistics)
9. Satchuthananthavale Rasiah Kukan Branavan, *Grounding Linguistic Analysis in Control Applications* (Massachusetts Institute of Technology, 2012) https://people.csail.mit.edu/branavan/papers/branavan-thesis.pdf
10. Ahead-of-time compilation, Wikipedia (2020) https://en.wikipedia.org/wiki/Ahead-of-time_compilation
11. Lemmatisation, Wikipedia (2020) https://en.wikipedia.org/wiki/Lemmatisation
12. K.V. Ramana, G.V. Ramesh, *Journal of Engineering Research and Application* 6(3) (2016) https://ru.scribd.com/document/320972667/Confiscation-of-Duplicate-Tuples-in-The-Relational-Databases
13. K. Wang et al., *NAACL-HLT* (2016) https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DEMO012.pdf
14. V.E. Meshkov, E.V. Meshkova, *Theses of the XIII International Scientific Conference Divnomorskoye*, 164-165 (2016)

15. K. Debbarma, B. Gopal Patra, D. Das, S. Bandyopadhyay, *3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)*, *COLING* (2012) https://www.researchgate.net/publication/255485610_Morphological_Analyzer_for_Kokborok