NLP and Attribution of Pseudonymic Texts:
Who Is Really the Author of the "Quiet Flows the Don"

Michail A. Marusenko (1), Rajmund H. Piotrowski (2), Yuri V. Romanov (2)

(1) Saint Petersburg State University, 12, 5th Sovetskaya str., apt. 4, St.Petersburg 193036, Russia
marusenko@mm1134.spb.edu
(2) Herzen State Pedagogical University of Russia, 48, Moyka Emb., St.Petersburg, 191186, Russia
romanov@YR4993.spb.edu

Abstract

The article is devoted to description of a newly developed linguistic automaton module for text attribution. Different methods for combining lexical, morphological and syntactical parameters are presented in their comparison. An evaluation of local design and functional techniques of the new module using the pattern recognition technique is presented in its application to the provocative checking of the challengeable authorship of the novel "Quiet Flows the Don" ascribed traditionally to the Nobel Prize winner Mikhail Sholokhov.

1. Introduction

The new developments in the NLP field, aimed at increasing of the efficiency of functions of a linguistic automaton (LA), insistently call for creating modules allowing to determine whether the processed document belongs to a certain thematic group or to a concrete author. Such modules, whose main function would be automatic sense extracting and text attribution, could also be used for solving criminal matters. The technique of these investigations is based on the principles of modern linguistic epistemology and can be presented as a scheme of a 5-link chain: original (linguistic object) - verbal hypothesis - theoretical formal model - reproducing model (NLP-module) - feedback (verification of conformity of the work of the model to the original functioning).

2. Text attribution

Taking into account the ideas of linguistic modeling, the attribution procedure can be represented as follows. For example, we have a certain text \( t_0 \), the authorship of which is not definitively established. Let’s assume now that we also have a set of texts \( T(A) = (t_1, t_2, \ldots, t_n) \), whose authorship is sufficiently well known of which we can be quite confident. The task of attribution is to compare \( t_0 \) with \( t_1, t_2, \ldots \), and to accept or reject the hypothesis that \( t_0 \) belongs to \( T(A) \). Any text attribution is of a probabilistic nature. To realize our attribution scheme (see above), one should use two kinds of technologies. The first is the analysis of the content level, the second deals with processing the expression level of the document under study.

Historical and philological analyses of a text are based on the theoretical postulates formulated in the works by K.Wentersdorf (1951: 161–193), V.V.Vinogradov (1961), Mcd.P.Jackson (1979). They enable the investigator to form an attribution hypothesis. The use of philological and extra-linguistic methods is a compulsory condition for realizing the procedure of attribution as it makes possible to put forward a verbal hypothesis which must be verified by means of formal technologies.

However, the semantic attribution method has two drawbacks: 1) it can't give any quantitative evaluation of reliability of it's results, 2) the real author of text \( t_0 \) can borrow the plot and the subject, as well as imitate the style and ideological orientation peculiar to the texts by other authors, though an investigation of the style-metrical parameters as well a paleographical and grapheological examination of text \( t_0 \) are also included. Unfortunately, this approach does not give any absolutely reliable results as the author can deliberately use the features characteristic of other individual authors. The more efficient methods are based on investigation of the linguo-statistical properties of the text. The most important among them are: 1) statistics of the sentence-length (Smith 1983: 77 - 75), 2) positional combinatorics of the most frequently-used words (Volochine 1995: 9 - 22), 3) statistics and combinatorics of rare words (Morton 1986: 1 - 8), 4) statistics of separate grammatical categories (Kjetsaa 1976: 5 - 24), 5) Markov's chains for solving text attribution (Khmelev 2000: 115 - 126). Though the use of a small number of simple indexes, and, particularly, of one linguo-statistical index-criterion, cannot guarantee a sufficiently reliable attribution of an anonymous text, such a restriction has allowed us to create a new standardized "a priori" set (AS) of 56 parameters, relevant to description of the structure and content of a sentence and a word-combination. The newly created attribution procedure is based upon a probability-matrix method employed in the pattern recognition theory (Ball 1969: 355). It involves five steps executed in a fixed sequence.
Step 1 consists in determining and description of an "a priori" alphabet of text classes (AA). For realizing this stage, the presence of, at least, two objects (texts $T_1$ and $T_2$) is necessary. These texts belong to, at least, two different, but already known, classes $\Omega_A$, $\Omega_B$... All classes in the AA are described in the language of the parameters of the AS.

Step 2 involves formation and standardization of the data matrix. The results of determination of the parameter values for the a priori classes (AC) and for the attributed objects can be represented in the form of a data matrix $Z$.

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Nj} & \cdots & z_{Nn} \end{bmatrix}.$$  

Here the number of objects is designated as $N$, and the number of parameters - as $n$. Then, element $z_{ij}$ shows the value which the $j$ parameter takes on the $i$ object.

The set of values of the $j$ parameter on all $N$ objects, shown in the $j$ column of the matrix, is designated as vector $z_j$. The $i$ line of the matrix, designated as $z_i$, shows the values taken by all the $n$ parameters on the $i$ object. Therefore, vector $z_j$ is the $N$-measured vector, and vector $z_i$ is the $n$-measured vector. Then, if the text (or the sample) undergoes examination in the volume of $N=100$ sentences, each of which is described by the set of $n=56$ parameters, then the total number of $z_{ij}$ is provided by $N \times n = 5600$ objects. In practice, when analyzing a text or a sample of a large volume, and some of the parameters are closely connected with the others, and the value of one parameter can be established basing on the values of another, then the opportunity arises to reduce the number of parameters used during the recognition procedure. As a result of such analysis, matrix $Z$ can be transformed into a more compact matrix $R = \{\mathbf{r}_{jk}\}_{j,k=1}^n$, the elements of which will be the $\mathbf{r}_{jk}$ correlation factors of the values of parameters $j$ and $k$. The smaller the size of $|\mathbf{r}_{jk}|$, the more difficult is the task to predict correctly the value of one parameter on the basis of the value of the other.

At Step 3 an algorithm of recognition is worked out. It presupposes a two-level procedure of recognition: deterministic (DAR) and probabilistic (PAR) ones.

Thus, for the DAR, if, on the one hand, basing on all informative parameters, the given object turns out to be attributed to class $\Omega_A$, and, on the other hand, if this object, basing on any informative parameters, does not turn out to be attributed to class $\Omega_B$, then it is related to $\Omega_A$ ($\Omega_B$, etc.). Classification of objects by the DAR is an iterative procedure. Each successive iteration causes a change in the contents and power of the text classes. Coordinates of the standards of classes are also re-computed.

The iteration of attribution is caused by the essence of the classification procedure. If the initial multitude is broken into $k$ classes, where $k$ is known a priori, the majority of the methods of classification operate by way of successive improvements of the initial division. The principal method of this group of procedures is the “method of mobile centers”.

If a “rejection” is received for all informative parameters for the given object, or if, at least, two parameters were found, on the basis of one of which the object relates to class $\Omega_A$, and upon the other, to class $\Omega_B$, then, in its response, the system will reject recognition. In this case, a PAR comes into use, which assigns probability or a degree of accurateness to each solution.

Step 4 is necessary to determine the volume of the sample. When determining the coordinates of the standards of classes and of the objects to be identified, two approaches are possible: 1) a complete run, or 2) a sampling test. Considering that the objects to be identified often include only a small text portion, to increase precision of the results when determining the coordinates of the objects it is expedient to conduct a full run of the text units. By determining the coordinates of the standards of classes, the volume of texts that make up the AC, allow to employ a sampling method.

Step 5 evaluates the quality of classification. The classifications, received as a result of our procedure, may be artifacts. Therefore, it is necessary to estimate the quality of the classifications. This estimation comes down to a search (for each class among the set of parameters of a sub-set), where the values of parameters will be the same for the majority of objects of the class being identified. These values will differ from the analogous values for other classes. In other words, the estimation of the quality of classification comes down to finding if such regularities cannot be determined, then the class is of no interest to the researcher.

For estimation of the quality of the object division into classes, we use criteria based on the measures of proximity (MP). For the MP between objects of one class $\Omega_p$, $p=1,2,\ldots,m$, an “average square dispersion” of objects is used.

The mathematical apparatus, described above, has been applied to attributing some more or less obscure texts. The most interesting results were obtained with respect to the Quiet Flows the Don - a masterpiece of the world literature written by the Nobel Prize winner Michail Sholokhov. Let us consider the basic stages of the procedure of its attribution.
Stage 1. There exist three hypotheses: 1) *Quiet Flows the Don* was written by M. Sholokhov ($H_0$), 2) it was written by one of the following pretenders: F. Kryukov, A. Serafimovich, S. Goloushev (Glagol) ($H_1$), 3) it is the result of cooperative work of several authors, including M. Sholokhov ($H^*_2$). In order to test the above-formulated hypotheses we build an AA of the classes (table 1).

In order to determine the coordinates of standards of the AC, a standard AA of parameters was used. From each corpus of texts represented in table 1, random samples in the volume of 200 sentences of the author’s text were taken, and the values of the parameters of the AA were determined.

Stage 2. Apportionment of parameters relative from the point of view of division of the AC. Statistical significance of the difference between the values of two parameters for a pair of authors (Sholokhov/Kryukov, Sholokhov/Serafimovich, Sholokhov/Glagol, Serafimovich/Kryukov, Serafimovich/Glagol, Kryukov/Glagol), was tested with the employment of the $t$-criteria of Student. Then, using the sub-set of style-differentiating parameters, a working-dictionary of five diagnostic parameters were formulated (table 2).

In correspondence with the procedure of pattern-recognition, the text was attributed in two steps: Stage 5 is executed with the help of the DAR and Stage 6 with the help of the PAR. The DAR determines the way the object being attributed belongs to one or other class of texts. As a deciding rule of the algorithm, the $t$-criteria of Student is used. Table 3 shows the result of the DAR work where the attribution of the original text of the novel was determined by 8.12%. The PAR attribution, with regards to which the DAR could not make a decision, is done upon the end of the DAR procedure. The PAR transforms the matrix of coordinates of the non-attributed text objects and AC into a matrix of suspended Euclid distances between objects and AC ($d(X_i, \Omega_j)$), which, in turn, are transformed into the matrix of probabilities of the objects belonging to AC ($P(X_i \in \Omega_j)$). The deciding rule is formulated here in the following way: $X_i \in \Omega_j \Leftrightarrow P(X_i \in \Omega_j) > 0.5$. In the given deciding rule the probability of the following objects belonging to AC is determined as follows: $\Omega_j$ (Sholokhov): td4813; $\Omega_j$ (Kryukov): td4711; $\Omega_j$ (Serafimovich): td2529, td4804. Here it would be necessary to state that “belonging” for the majority of the objects was not established.

<table>
<thead>
<tr>
<th>Parameter Code</th>
<th>Name of the parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>X14</td>
<td>The number (n.) of subordinate clauses without a conjugated form of the verb in a uniform sentence (US).</td>
</tr>
<tr>
<td>X18</td>
<td>The number of connective word forms (w/f's) in a US</td>
</tr>
<tr>
<td>X21</td>
<td>The n. of pronouns in a US</td>
</tr>
<tr>
<td>X27</td>
<td>The n. of subordinate conjunctions in a uniform sentence</td>
</tr>
<tr>
<td>X30</td>
<td>The n. of words in the accusative in a uniform sentence</td>
</tr>
</tbody>
</table>

Table 1.

<table>
<thead>
<tr>
<th>Code</th>
<th>Author, title, publication data</th>
<th>Sentences (N=)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>Глагол С. На юг. 19(?)</td>
<td>1300</td>
</tr>
<tr>
<td>S01</td>
<td>Серафимович А. Бабки деревня и др. рассказы. 1923</td>
<td>400</td>
</tr>
<tr>
<td>S02</td>
<td>Серафимович А.С. Железный поток. 1924</td>
<td>2000</td>
</tr>
<tr>
<td>K01</td>
<td>Крюков Ф.Д. Рассказы. 1914</td>
<td>6650</td>
</tr>
<tr>
<td>Sh01</td>
<td>Шолохов М.А. Донские рассказы. 1926</td>
<td>1500</td>
</tr>
</tbody>
</table>

On Stage 3, it’s necessary to complete the procedure of transition to mathematical models of AC, where each class of texts is presented in the form of coordinates of a standard of the given class. Since we work with big bulks of texts, it would be more logical to employ the sampling method.

Stage 4, realizing the transition from a real object (the text undergoing the procedure of attribution) to its mathematical model, consists in describing this object in the language of parameters of the working-dictionary of the pattern recognition system. If one has a hypothesis about the existence of the author and co-author, then it is not necessary to use the whole novel as an object of attribution. In such a case the mathematical model would include properties of at least two known qualitatively dissimilar sum-totals (of individual styles of the different writers). Therefore, as a unit of attribution we take not the whole text of the novel but rather each individual chapter of *Quiet Flows the Don*. Altogether, there are 228 chapters in the novel. The total volume of speech in the author’s voice is 16,796 sentences.

Each chapter has its own code, for which the combination of “td” designates belonging to the *Quiet Flows the Don*. E.g. the code td2512 is a mark that we are dealing with the 2nd volume, the 5th section, the 12th chapter of Sholokhov’s novel.
Table 3

<table>
<thead>
<tr>
<th>Class</th>
<th>Make-up of the class</th>
<th>Rating</th>
<th>Volume</th>
<th>Volume as a % from 16,796</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ω (Sholokhov)</td>
<td>td1215, td2509, td3621, td3643, td4704, td4710, td4719</td>
<td>7</td>
<td>269</td>
<td>1.60</td>
</tr>
<tr>
<td>Ω (Kryukov)</td>
<td>td4722, td4807, td4809</td>
<td>3</td>
<td>184</td>
<td>1.09</td>
</tr>
<tr>
<td>Ω (Serafimovich)</td>
<td>td1101, td1107, td1112, td1318, td2517, td2520, td2526, td3604, td3651, td4709, td4716, td4717, td4726, td4812</td>
<td>14</td>
<td>877</td>
<td>5.22</td>
</tr>
<tr>
<td>Ω (Glagol)</td>
<td>td4723</td>
<td>1</td>
<td>35</td>
<td>0.21</td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td>25</td>
<td>1365</td>
</tr>
</tbody>
</table>

However, the analysis of statistical data allows to display some regularities in the distribution of probabilities of objects belonging to AC: thus, in turn, it is striking that the sum of probabilities of objects manifests itself for the classes

Ω₁ (Sholokhov)& Ω₂ (Kryukov),
Ω₃ (Kryukov)& Ω₄ (Serafimovich),
Ω₅ (Serafimovich)& Ω₆ (Glagol). This allows one to suspect the existence of several a posteriori classes, each of which includes texts belonging to the pen of two authors at a time: Ω₅ (Sholokhov, Serafimovich),
Ω₆ (Sholokhov, Kryukov), Ω₇ (Serafimovich, Kryukov), Ω₈ (Serafimovich, Glagol).

The results of PAR relative to the four a priori and the four a posteriori classes allowed us to determine to which author the 91.88% of text in the author’s voice belongs. The obvious result would be the fact that M. Sholokhov is not the only author of the text of the novel *Quiet Flows the Don*, and, furthermore, is not even one of the main authors of the novel. Only 2.39% of the novel text is attributed to Sholokhov as an individual author. First comes A. Serafimovich, to whom 5.91% of the text is attributed. In total, 1.37% and 0.21% of text in the author’s voice are fragments of texts by F. Kryukov and S. Goloushev respectively. The process of creation of the remaining text of the novel can be represented as follows: 63.68% of the text carries imprints of the writing manner of the two authors, M. Sholokhov and A. Serafimovich. The rest of the text (about 27%) is the result of “completion” by A. Serafimovich of texts by F. Kryukov and S. Goloushev. Thus, the genuine author of the novel *Quiet Flows the Don* could be identified as A. Serafimovich.

There is one circumstance, however, which does not allow one to complete our procedure at this point. The problem is that when forming the original hypothesis (*H₈*), some biographical and ideological principles were used. The comparison of the text of the novel *Quiet Flows the Don* with M. Sholokhov’s texts of "Донские рассказы" (*The Don Stories*), which are close to *Quiet Flows the Don* both chronologically and thematically, shows serious ideological differences belonging to classes Ω₁ (Sholokhov) and Ω₇ (Serafimovich) regularly exceeds the threshold value of 0.5; a similar normality existing between these two works. As a result, the disparity in the genre of the two texts fell from attention, even though stylistic studies persuasively proved that the differences in the genre between the works of the same author are often deeper than the differences between the authors working within the same genre. Therefore, it would be expedient to correct the initial hypothesis, adding to it one more AC, designated as Ω(Glagol and Sholokhov). The rest of the text includes fragments of texts by F. Kryukov and S. Goloushev.

In accordance with the concept of pattern recognition, if correction of the hypothesis of attribution leads to a change in the make-up of the AC, then the whole procedure of classification, using the DAR- and PAR-techniques, must be repeated. The analysis of the results of the new procedure allows us to make the following conclusions:

1. If we begin with the premise that one author working alone wrote *Quiet Flows the Don*, then, without a doubt, most likely, it is Sholokhov with 26 chapters and 8.59% of the text in the author’s voice.

2. The majority of text in the author’s voice demonstrates proximity to the standards of classes of pairs of authors, while the main participant in such “collective authorship” is A. Serafimovich: 191 chapters or 86.99% of the text in the author’s voice in “cooperation” with other authors.

3. Sholokhov “cooperated” with A. Serafimovich (140 chapters or 65.19% of the text), and partly with F. Kryukov (1 chapter or 0.16% of the text). Hence the most probable hypothesis is that the genuine author of *Quiet Flows the Don* was A. Serafimovich, who wrote the novel in co-authorship with M. Sholokhov (or using his texts), or also using texts of F. Kryukov and S. Goloushev (Glagol).

3. Conclusion

In order to complete our modeling experiment, it’s necessary that some biographical and ideological explanation should be presented here. But it might
astonish a lot of readers and experts on Soviet Literature of the XXth century. Let’s deal with the version of the authorship of *Quiet Flows the Don* which was once suggested by the literary critic М.А.Аникин (1993). Prof. Аникин believes that A. Серафимович (1863-1949) was the most prominent pre-Revolutionary writer of the Don, the most first-rate expert of the region’s everyday life and customs, who cultivated the Don theme even before the Revolution of 1917. He was a refined member of the intelligentsia, received a diploma of the St. Petersburg University, and the author of genuinely interesting novels. He had a very rich biography, having fought in both the First and Second World Wars. The main reason why the writer took on a literary mystification is that *Quiet Flows the Don* was a very risky work for a period of proletarian literature. For this reason the food-requisitioner Шолохов with four years of school education, a person of a stainless communist reputation was for Серафимович a reliable shield. He gave him an opportunity to write without being prosecuted. It was Серафимович who noticed Шолохов, and wrote an introduction to "Донские рассказы" (*The Don Stories*). He gave an impulse to the literary “development” of Шолохов. After the death of Серафимович the literary talent of Шолохов began to die down.

**Abbreviations**

AA - *a priori* alphabet  
AC - *a priori* classes  
AS - *a priori* set  
DAR - deterministic algorithm of recognition  
PAR - probabilistic algorithm of recognition

4. References


**Acknowledgement**

The present research is sponsored by the Russian Fund of Fundamental Investigations (RGNF), project No. 02-04-00195а.