electronics

MDPI

Article

Software-Based Approach Towards Automated Authorship Acknowledgement—Chi-Square Test on One Consonant Group

Iryna Khomytska [1], Vasyl Teslyuk [2], Natalia Kryvinska [3,4,*] and Iryna Bazylevych [5]

[1] Applied Linguistics Department, Lviv Polytechnic National University, 79013 Lviv, Ukraine; iryna.khomytska@ukr.net
[2] Computer Aided Design Systems Department, Lviv Polytechnic National University, 79013 Lviv, Ukraine; vasyl.m.teslyuk@lpnu.ua
[3] Department of e-Business, School of Business, Economics and Statistics, University of Vienna, A-1090 Vienna, Austria
[4] Department of Information Systems, Faculty of Management, Comenius University in Bratislava, 25 82005 Bratislava, Slovakia
[5] Theoretical and Applied Statistics Department, Ivan Franko National University of Lviv, 79000 Lviv, Ukraine; i_bazylevych@yahoo.com
* Correspondence: natalia.kryvinska@univie.ac.at

Received: 22 June 2020; Accepted: 9 July 2020; Published: 13 July 2020

Abstract: A one-consonant group approach to the authorship attribution has been proposed. The approach is based on determining, by the chi-square test, the consonant group in which the difference between the texts by different authors is statistically significant. The developed model determines author-differentiating capability of each consonant group in a relation of the number of comparisons, in which the difference between the texts by two authors is statistically significant to the total number of comparisons. The determined general author-differentiating capability of the group of stop consonants, which is a statistical parameter of the authorial style, is the highest in the comparisons of texts from the publicist and belles-lettres styles. The one-consonant group approach simplifies the whole process of authorship attribution and ensures a higher level of automation. The conducted experiments on the Java programming language have proved that the chi-square test is a powerful nonparametric statistical test that can be used for author identification on the level of English consonants with a test validity of 95%.

Keywords: consonant group; chi-square test; nonparametric statistical test; author-differentiating capability; java-based model; software framework

1. Introduction

A language is not a strictly arranged system and has probabilistic and stochastic character. In this case, it is advisable to apply the statistical methods. The analysis of recent publications has shown that authorship attribution has been performed on all language levels: morphological, lexical, and syntactical [1,2]. The choice of a language level is of great importance for formalization as the level structure must be strict. With regard to the lexical level, its structure is not easy to formalize because of neologisms and foreign loans that constantly enlarge the vocabulary. Moreover, polysemantic words, lexical stylistic devices, and idioms may cause ambiguous interpreting [3,4]. Syntactical structures vary from simple to complicated. The latter are difficult to formalize [5–7]. The advantage of our research is the choice of the phonological level. Unlike the other language levels, on this level, the number of elements (phonemes) is unchangeable and consequently, the level has

more strict structure. Regarding the statistical methods applied for authorship attribution, their success depends on the chosen language level and on the given sample. To ensure more reliable results, it is recommended to use a combination of statistical methods most suitable for the research [8]. It must be noted that the nonparametric methods are easier to use as it is not necessary to prove that the sample is distributed according to the normal distribution law [9–11]. The analysis of recent research of the authorship attribution performed by the chi-square test has shown that it was done "…to assess the significance of the measured difference between the 'sample' and the 'reference' texts." The most plausible author was established by the revealed statistically significant differences in word lengths, sentence lengths, paragraph lengths, in the frequency of occurrence of letters, and punctuation marks. The test validity was 90%. The results were applied in forensic authorship identification. Such study is of great importance as it serves the cause of justice [12]. According to the results of analysis of the state-of-the-art software developed for author identification, the most common programming languages are: R, Python, and Java [13–15]. Our research is done on the Java programming language. The strength of the chi-square test was evaluated in our previous research in a comparison with the Student's t-test, the ranking method, and the style distance determination method. The analysis of the obtained results has shown that for authorship attribution on the level of consonant groups, the chi-square test is more powerful than the Student's t-test, the ranking method, and the style distance determination method. On average, the texts by different authors differ essentially in 6 of 8 consonant groups if the chi-square test is applied, and in 4–6 of 8 groups if the other three mentioned methods are applied [16]. In addition to this, the chi-square test is a nonparametric test for which the samples compared should not follow the law of normal distribution. Because of this, the test is easier to use than the parametric tests and considerably simplifies author identification. The efficacy of the chi-square test in this research has been proved on the material of presidential speeches by B. Obama and D. Trump, the newspaper articles by D. Webster and S. Logan, and the pieces of English emotive prose by E. Bronte. The speaking pattern of the speaker in the formal speech can be identified if compared with a number of texts by one and the same speaker. To ensure high test validity, the phonological language level has been chosen. The elements of this level are phonemes. By phonemes, here, transcription symbols of consonants and vowels are meant. The research is focused on the consonants. The texts by different authors are differentiated by consonants grouped according to the following classification: labial, dorsal, coronal, fricative, nasal, sonorous, velar, and stop consonants. The consonant groups of labials, dorsals, velars, and fricatives were researched in the previous papers [8,17]. To simplify the author identification process, it is proposed to differentiate texts in one of eight consonant groups. The group must have a high author-differentiating capability. To determine the group-differentiating capability, it is necessary to make a sufficient number of experiments in which the text under study by one author is compared with another one by another author. The texts compared must be of the same style, genre, and topic. The mentioned three factors should be taken into account to make the individual author's characteristics vivid and easier to detect. Otherwise, the overlapping of the three factors is sure to bring about difficulties in interpreting the differences caused by another style, genre, topic, and author's writing peculiarities. If the requirements to homogeneity of the texts are met, the results obtained are robust. The consonant group-differentiating capability is a parameter of general authorial characteristics, which is determined by the developed model. The consonant group in which the statistically significant differences are obtained in all or nearly all comparisons is considered to have high differentiating power. The purpose of the research is to ensure simpler and more automated authorship attribution. The simplification is done by reducing the number of consonant groups to one. The results of the research can be applied for author identification in anonymous e-mail communication that includes malware, phishing and spamming, in forensic and scientific fields, as well as in the belles-lettres style.

2. Mathematical Support of Software System

2.1. The Method Developed

To determine author-differentiating capability of each consonant group, a powerful nonparametric chi-square test has been used. One group with high author-differentiating capability is chosen for author identification. The high author-differentiating capability of a consonant group is established in the case of obtained statistically significant differences in a sufficient number of comparisons of texts by different authors.

The developed algorithm of author identification is based on application of the chi-square test.

1. Two texts from the newspaper articles by D. Webster and S. Logan chosen for disputed authorship are transcribed.

2. The consonants are singled out from the bulk of consonants and vowels.

3. The sample of consonants (51,000) is divided into portions (1000 each).

4. The number of consonants in each portion is calculated.

5. The consonants are united in 8 consonant groups. The classification is the following;

(a) according to the place of obstruction: labial: p, b, m, f, v, w; dorsal: θ, ð, t, tʃ, dʒ, ʃ, ʒ, r; coronal: t, d, n, s, z, l, velar: k, g, ŋ;

(b) according to the manner of production of noise and according to the type of obstruction: nasal: m, n, ŋ; sonorous: w, r, j, l; fricative: f, v, θ, ð, h, s, z; stop: p, b, t, d, k, g, tʃ, dʒ.

6. The number of consonants in groups is calculated.

7. The data are divided into intervals: the number of intervals s is determined by: $s = \sqrt{n}$ [9], where n is a number of portions in a sample. The width of the interval h is determined by: $h = \frac{\max - \min}{s}$, where $\max$ is a maximum value, min is a minimum value [18–20].

8. The number of frequencies ($\vartheta_{i,j}$) getting into the i-th interval of the 1-st sample j and the 2-nd sample ($\vartheta_{i,j}$) is calculated.

9. The general relative frequency of getting into the i-th interval of the two samples: $\frac{n_j \vartheta_i}{n}$ is determined. The sample size (a number of portions) of the 1-st sample j is n_j. The size of two samples is n. The random variable is ϑ_i.

10. The level of significance is 5%.

11. The statistic $\hat{\chi}_n^2$ is determined by the difference of the number of frequencies getting into the i-th interval of j-th text and the general relative frequency of getting into the i-th interval of two samples (ϑ_i).

$$\hat{\chi}_n^2 = \sum_{i=1}^{s} \sum_{j=1}^{k} \frac{\left(\vartheta_{i,j} - \frac{n_j \vartheta_j}{n} \right)^2}{\frac{n_j \vartheta_j}{n}}, \quad \vartheta_j = \sum_{j=1}^{k} \vartheta_{ij} . \tag{1}$$

The texts by two authors can be differentiated if $\hat{\chi}_n^2 \geq \chi^2_{1-\alpha,(s-1)(k-1)}$, where α is a significance level, $(s-1)(k-1)$ is a number of degrees of freedom, s is a number of intervals, k is a number of texts [21–23].

12. The author-differentiating capability of the consonant groups is established by the chi-square test.

13. The consonant group with high capability of author differentiation is identified.

14. The text authors D. Webster and S. Logan are considered significantly distinct if the established difference is statistically significant.

15 The same procedure is done in pairwise comparisons of texts by B. Obama, D. Trump, and E. Bronte.

16. The process of authorship attribution is made simpler and more automated by performing it in one consonant group.

The advantage of the developed algorithm is simplicity and automation. The author can be identified only in one consonant group in which the samples differ essentially.

2.2. The Developed Model

A model for determining the author-differentiating capability of consonant groups (ADC) has been developed. The model determines the authorial style characteristics in a relation to the number of cases in which statistically significant differences were established between the texts by different authors in the given consonant group (*SSD*) to the total number of cases of comparisons of the texts by different authors in the given consonant group (*TNC*):

$$ADC = \frac{SSD}{TNC}. \quad (2)$$

To establish the author-differentiating capability of consonant groups, seven pairwise comparisons were analyzed. Such an approach is supposed to help obtain more reliable information about the authorial style characteristics in each consonant group as more vocabulary is covered and specifics of consonant group functioning get clear cut. Consequently, for 7 comparisons, statistically significant differences (SSD) can be revealed: $0 \le SSD \le 7$. The consonant group in which statistically significant differences are obtained in 6 or 7 comparisons is considered to have the highest author-differentiating capability. This consonant group is chosen to perform author identification. The developed model for one consonant group ensures more simplified and automated authorship attribution.

3. The Developed Software

The developed software on the Java programming language for author identification is aimed at simplifying the whole process with a test validity of 95%. The simplification involves the reduction of the number of consonant groups in which the statistically significant differences between the texts by two authors have been revealed. The powerful statistical test—the chi-square—test has been used to identify the consonant group that can be used alone instead of eight groups. The structure of the developed program system consists of six main modules, user interface, data base, and links with libraries (Figure 1). The modular arrangement of the program allows us to improve the program product fast. The Consonant Sample Formation Module is for removing all vowels and punctuation marks from the sample. The next modules are for interval division, calculation of consonants in portions and groups, and doing the chi-square test for text differentiation. The algorithm of identifying the group with a high differentiating capability is presented in Figure 2. An example of the main menu of the program system is given in Figure 3.

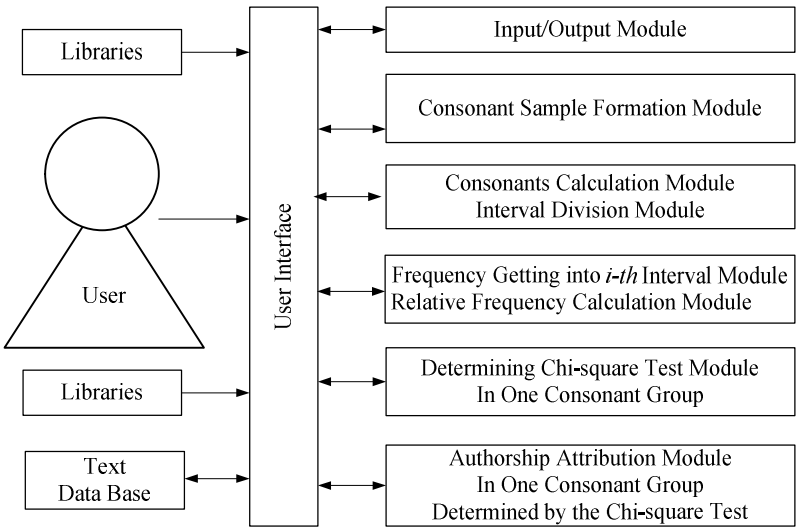

Figure 1. The structure of the developed program system.

The structure of classes of the developed software is as follows: "Text", "Transcribed Text", "Consonant Sample", "Sample Division into Portions", "Sample Division into Groups"—the consonants are united into eight consonant groups according to their acoustic-articulatory features: "Calculating Consonants", "Interval Division", "Calculating the Number of Frequencies Getting into the i-th Interval", "Performing Chi-square test", "Identifying Powerful Consonant Group", "Text Difference in Consonant Groups". The software classes responsible for the statistical processing of the sample are presented in the diagram in Figure 4. As the Java programming language has been used, the developed software is platform-independent [24,25]. In the developed program for information, ensure the list ArrayList<String> is used. It is an automated expanding list [26,27]. The list is good for performing operations with dynamic data. An example of the used data list structure is given in Figure 5.

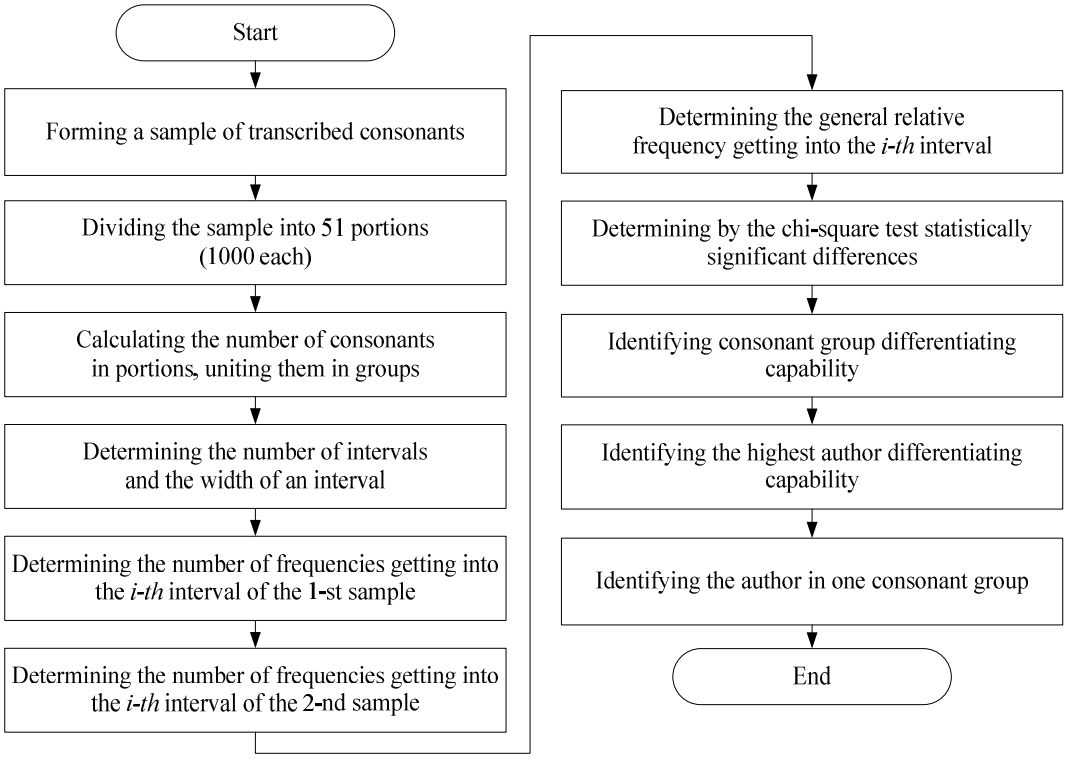

Figure 2. The algorithm of identifying the group with a high differentiating capability.

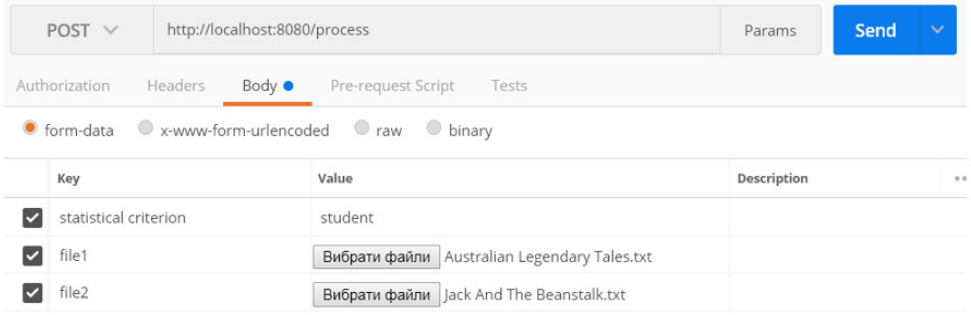

Figure 3. An example of the main menu of the program.

The list structures make it possible to compare a word from a new text with already-existing words in the program. If the word is in the program, there is the command in the program that the word with index "x" needs its transcription variant. In this case, the list gives the transcribed equivalent of the word. If the word is not in the program, the symbol is added to the list with the index equal to the index of its transcription in the list.

The developed software identifies the author in one consonant group. In this research the group of stop consonants has been used as the strongest in text differentiation. The determined consonant group can be used for further text differentiation by the same authors. The in-built data base H2 has a sufficient amount of transcription symbols and makes the software system in most cases independent of the transcription site. The sample size necessary for obtaining results with a test validity of 95% is 50,000 consonants.

4. Results of the Study

The experiments have been aimed at determining the consonant group in which the author can be identified. The group has been established by the chi-square test. This is the group of stop consonants. The group has the following consonants: p, b, t, d, k, g, tʃ, dʒ. The group has higher differentiating capability than the other consonant groups considered in this paper: the nasal, sonorous, and coronal consonant groups. The whole number of consonant groups is eight, four of which (labial, dorsal, velar, and fricative) were researched in the previous papers [8,16,17]. For the material of research, the texts from the publicist style have been chosen. These are the texts from presidential speeches by B. Obama and D. Trump, the newspaper articles by D. Webster and S. Logan, and the pieces of English emotive prose by E. Bronte. The comparison of two samples by Bronte is aimed at revealing the difference between different chapters by the same author, which characterizes efficacy of the chi-square test. Such an approach has been applied to make a thorough analysis of individual authorial styles in the group of stop consonants. The results have shown that, in this consonant group, statistically significant differences have been obtained in all but one comparison. The other consonant groups analyzed in this research—nasal, sonorous, and coronal—have lower author-differentiating capability. This proves that the stop consonant group has the highest author-differentiating capability and can be taken alone to identify the author. The authorial style has been identified in all conducted comparisons with a test validity of 95%. The results found in the groups of stop, nasal, sonorous, and coronal consonants are given in Tables 1–4.

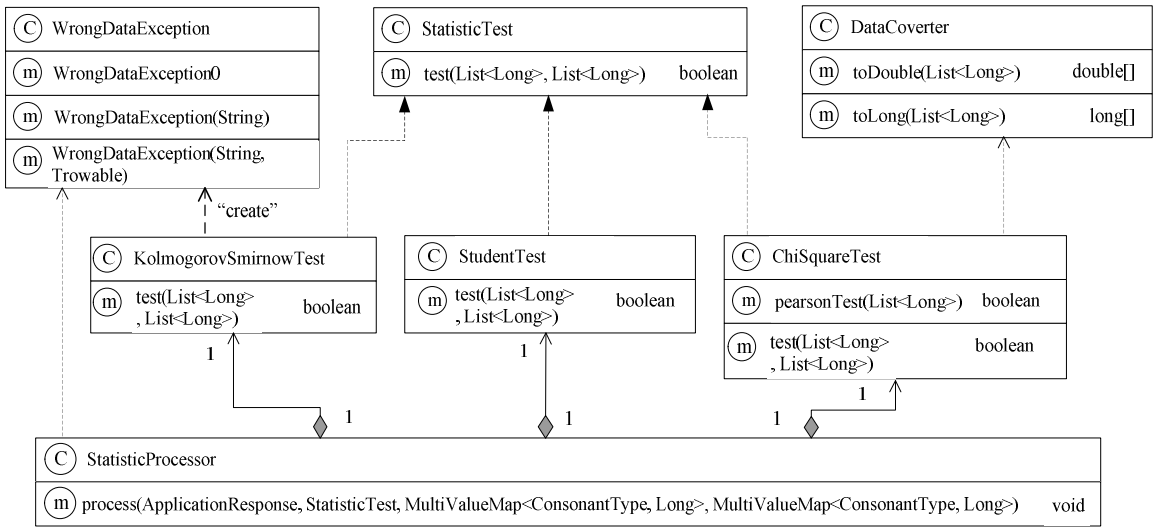

Figure 4. A diagram of the system classes for the statistical processing of a sample.

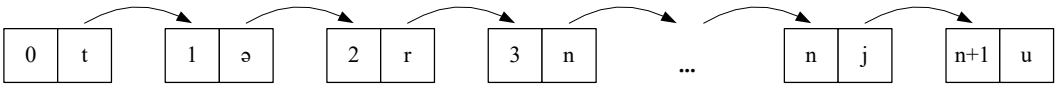

Figure 5. An example of the used data list structure.

The statistically significant differences have been revealed in the stop phoneme group in six of seven comparisons (Table 1). In the comparison of Trump-Logan, the difference is unessential because in the presidential speeches by D. Trump and newspaper articles by S. Logan, similar political and legal issues are discussed. The samples have some vocabulary in common. Another reason for this result is the phonological specifics of stop consonants functioning.

Table 1. The author-differentiating capability of the stop phoneme group.

Compared Texts by Different Authors	Author-Differentiatng Capability
Obama-Trump	+
Obama-Webster	+
Obama-Logan	+
Trump-Webster	+
Trump-Logan	-
Webster-Logan	+
Bronte-Bronte	+

In Table 2, the results for the nasal consonant group are presented. The author-differentiating capability of the group of nasal consonants has been established in five of seven comparisons. The nasal consonant group is the second one in which the unessential difference has been obtained in the comparison of Trump-Logan. Similarity (unessential difference) on the phonological level is observed in another comparison: Obama-Trump. As in both samples, political issues are discussed, and the lexical similarity corresponds to the phonological one.

Table 2. The author-differentiating capability of the nasal phoneme group.

Compared Texts by Different Authors	Author-Differentiatng Capability
Obama-Trump	-
Obama-Webster	+
Obama-Logan	+
Trump-Webster	+
Trump-Logan	-
Webster-Logan	+
Bronte-Bronte	+

The common social-political issues in the samples by Obama and Logan, as well as similar vocabulary from two chapters of emotive prose by Bronte, have brought about unessential differences in both comparisons in the sonorous consonant group. In the other comparisons, the statistically significant differences have been established (Table 3).

Table 3. The author-differentiating capability of the sonorous phoneme group.

Compared Texts by Different Authors	Author-Differentiating Capability
Obama-Trump	+
Obama-Webster	+
Obama-Logan	-
Trump-Webster	+
Trump-Logan	+
Webster-Logan	+
Bronte-Bronte	-

The least author-differentiating capability has been established in the group of coronals. The samples have statistically significant difference in four of seven comparisons. It is evident that the data obtained on the phonological level reflect both the peculiarities of the lexical level and inner phonological laws (Table 4).

Table 4. The author-differentiating capability of the coronal phoneme group.

Compared Texts by Different Authors	Author-Differentiating Capability
Obama-Trump	-
Obama-Webster	+
Obama-Logan	+
Trump-Webster	-
Trump-Logan	+
Webster-Logan	-
Bronte-Bronte	+

The text differentiation analysis has shown that the author-differentiating capability and topic-differentiating capability of the stop consonants is the highest in a comparison with the nasal, sonorous, and coronal groups and the researched samples can be differentiated in this sole group. The author-differentiating capability of the stop, nasal, sonorous, and coronal consonant groups obtained by the Equation (2) is given in Table 5.

Table 5. The author-differentiating capability of consonant groups.

Consonant Group	Author Differentiating Capability (ADC)
stop	0.86
nasal	0.71
sonorous	0.71
coronal	0.57

The data obtained for the stop consonant group can be used to perform author identification and topic identification in the texts of the publicist style and the emotive prose of similar topic by the researched authors.

5. Discussion

The developed model for determining author-differentiating capability of consonant groups has made it possible to conclude that the stop consonant group has the highest author-differentiating capability: ADC = 0.86. The other researched groups have the following general differentiating capability: for the nasal and sonorous groups, ADC = 0.71, for the coronal group, ADC = 0.57. The results obtained are valid for the samples from presidential speeches by B. Obama and D. Trump, the newspaper articles by D. Webster and S. Logan, the pieces of English emotive prose by E. Bronte, and the publicist style in general. The experiment with the texts by one author (E. Bronte), but on different topics, has proved a possibility of topic identification. The test validity of the results equals 95%. The highest author-differentiating capability of the stop consonant group has made it possible to apply the proposed one-consonant group approach for author identification which is simpler and more automated. Considering limitations of the chi-square test, it should be noted that it is very sensitive to the sample size. The conducted experiments have shown that the sample size of 51 portions with 1000 consonant phonemes in each portion is sufficient for obtaining reliable data. The second limitation deals with the language level, on which the research is done. The test validity is higher on the phonological level (95%) [8,16,17] than on the lexical level (90%) [12]. Specifics of each consonant group are related to one more limitation. This research has revealed that the stop consonant group has the highest author-differentiating capability if the chi-square test is applied. The other consonant groups do not give the same results. Consequently, the chi-square test shows its efficacy in definite consonant phoneme groups.

6. Conclusions

One of the advantages of the research is the phonological level on which it is done. Because of the level's strict arrangement, it is the easiest for formalization. The one-consonant group approach for authorship attribution has been proposed. The chi-square test-based model has been developed for determining the consonant group in which the author can be identified. The model determines the author-differentiating capability of each consonant group in a relation to the number of cases in which the statistically significant difference is established between the compared authors to the total number of comparisons. The identification of an author is done in the four following groups of consonants: the stop consonants, the nasal consonant, the sonorous consonants, and the coronal consonants. The obtained data have shown that the stop consonant group has the highest author-differentiating capability whose ADC = 0.86. In this group, the texts differ essentially in the following comparisons: Obama-Trump, Obama-Webster, Obama-Logan, Trump-Webster, Webster-Logan, and Bronte-Bronte. In the last comparison, the topic-differentiating capability has been revealed in the text by E. Bronte. The last result is important as it characterizes the specifics of the authorial style. The applied approach has proved that it is possible to differentiate authors in one consonant group. The advantage of applying the nonparametric chi-square test is its simplicity if compared with the parametric tests. The chi-square test has proved efficient for authorship attribution and topic attribution (Bronte-Bronte) on the level of consonant groups. The data have been obtained with a test validity of 95%. The developed algorithm for author identification has been implemented on the Java programming language. Its choice is well grounded as it is platform-independent and the modular principle of the structure of the software system allows us to quickly modify and improve the program. As a higher level of automation requires reduction of the number of consonant groups, in which the author is identified, the software system is simpler and more automated. The future research will focus on testing the other statistical tests for authorship attribution on the phonological level. The parametric and nonparametric tests can be tested separately and in different combinations. The advantage of combinations of tests is that the results

obtained by one test can be verified by the other tests. If the same results are obtained by more than one test, the data are more reliable.

7. Prospects for Future Research

The future research will focus on testing the other statistical tests for authorship attribution on the phonological level. The parametric and nonparametric tests can be tested separately and in different combinations. The advantage of the use of combinations of tests is that the results obtained by one test can be verified by the other tests. If the same results are obtained by more than one test, the data are more reliable.

Author Contributions: Formal analysis, V.T. and I.K.; investigation, V.T. and I.K.; method, I.K. and I.B.; model, I.K. and V.T.; resources, N.K. and I.B.; software, V.T. and I.K.; validation, V.T. and I.B.; writing—original draft preparation, I.K. and V.T.; writing—review and editing, V.T. and N.K.; visualization, V.T. and I.K.; supervision, V.T.; data curation, I.K. and N.K. All authors have read and agreed to the published version of the manuscript.

Funding: Open Access Funding by the University of Vienna.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bevendorff, J.; Ghanem, B.; Giachanou, A.; Kestemont, M.; Manjavacas, E.; Potthast, M.; Rangel, F.; Rosso, P.; Specht, G.; Stamatatos, E.; Stein, B.; Wiegmann, M.; Zangerle, E. Shared Tasks on Authorship Analysis at PAN 2020. In Proceedings of the European Conference on Information Retrieval, Lisbon, Portugal, 14–17 April 2020; pp. 508–516, doi:10.1007/978-3-030-45442-5_66.
2. Lytvyn, V. Development of a method for the recognition of author's style in the ukrainian language texts based on linguometry, stylemetry and glottochronology. *East. -Eur. J. Enterp. Technol.* **2017**, 4, 10–18.
3. Vysotska, V.; Burov, Ye.; Lytvyn, V.; Demchuk, A. Defining Author's Style for Plagiarism Detection in Academic Environment. In Proceedings of the IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 128–133, doi:10.1109/DSMP.2018.8478574.
4. Tamboli, M. S.; Prasad, R. S. Authorship Identification with Multi Sequence Word Selection Method. In: Abraham A., Cherukuri A., Melin P., Gandhi N., Eds.; Intelligent Systems Design and Applications. ISDA 2018. *Adv. Intell. Syst. Comput.* **2020**, *940*, 653–661, doi:10.1007/978-3-030-16657-1_61.
5. Bisikalo, O. V. Sentence syntactic analysis application to keywords identification ukrainian texts. *Radio Electron. Comput. Sci. Control.* **2016**, 3, 54—65.
6. Bhargava, M.; Mehndiratta, P.; Asawa, K. Stylometric Analysis for Authorship Attribution on Twitter. In Proceedings of the Second International Conference on Big Data Analytics, Mysore, India, 16–18 December 2013; pp. 37–47, doi:10.1007/978-3-319-03689-2_3.
7. Bozkurt, I., N.; Baghoglu, O.; Uyar, E. Authorship attribution. In Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS), Ankara, Turkey, 7–9 November 2007; pp. 1–5, doi:10.1109/ISCIS.2007.4456854.
8. Khomytska, I.; Teslyuk, V.; Kryvinska, N.; Beregovskyi, V. The Nonparametric Method for Differentiation of Phonostatistical Structures of Authorial Style. *Procedia Comput. Science* **2019**, *160*, 38–45.
9. Koppel, M.; Schler, J.; Argamon, Sh. Authorship Attribution: What's Easy and What's Hard? *SSRN Electron. J.* **2013**, 21, 317–331, doi:10.2139/ssrn.2274891.
10. Azarbonyad, H.; Dehghani, M.; Marx, M.; Kamps, J. Time-Aware Authorship Attribution for Short Text Streams. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 727–730, doi:10.1145/2766462.2767799.
11. Jamak, A.; Alen, S.; Can, M. Principal Component Analysis for Authorship Attribution. *Bus. Syst. Res.* **2012**, *3*, 49–56.
12. Guillén, Nieto, V.; Vargas, Sierra, C.; Pardiño, Juan, M.; Martínez, Barco, P.; Suárez, Cueto, A. Exploring State-of-the-Art Software for Forensic Authorship Identification. *Int. J. Engl. Stud.* **2008**, *8*, 1–28.
13. Schmid, M. R.; Farkhund Iqbal, B.; Fung, C.M. E-mail authorship attribution using customized associative classification. *DFRWS* **2015**, *14*, S116–S126.

14. Argamon, Sh.; Koppel, M.; Pennebaker, J.; Schler, J. Automatically Profiling the Author of an Anonymous Text. *Commun. ACM* **2009**, 52, 119–123, doi:10.1145/1461928.1461959
15. Juala, P. Authorship Attribution. *Found. Trends® Inf. Retr.* **2008**, 1, 233–334, doi:10.1561/1500000005.
16. Khomytska, I.; Teslyuk, V. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. *In Advances in Intelligent Systems and Computing*; Shakhovska, N., Ed.; Springer: Lviv, Ukraine, 2016; Volume 512, pp. 149–163.
17. Khomytska, I.; Teslyuk, V. Statistical Models for Authorship Attribution. *In Advances in Intelligent Systems and Computing*; Shakhovska, N., Medykovskyy, M., Eds.; Springer: Lviv, Ukraine, 2019; Volume 1080, pp. 579–592.
18. Watanabe, S. *Probability Theory and Mathematical Statistics*; Springer: Kyoto, Japan, 1988.
19. Gries, Th. S. *Statistics for Linguistics with R: A Practical Introduction (Trends in Linguistics: Studies & Monographs)*; Mouton de Gruyter: Berlin, Germany, 2009; p. 348.
20. Rozanov, Iu. A.; Silverman, R. A. *Probability Theory: A Concise Course*; Dover Publications Inc: New York, NY, USA, 1977.
21. Jorgensen, P.E.T. *Analysis and Probability: Wavelets, Signals, Fractals*; Springer Science + Business Media LLC: New York, NY, USA, 2006.
22. Bhattacharya, Rabi; Waymire, Edward C. A *Basic Course in Probability Theory*; Springer, 2nd ed.; Springer: Cham, Switzerland, 2016. ISBN 978-3-319-47974-3.
23. Everitt, B.S. *Cambridge Dictionary of Statistics*; Cambridge University Press: Cambridge, UK, 1998.
24. Kaczor, S.; Kryvinska, N. It is all about Services - Fundamentals, Drivers, and Business Models, *J. Serv. Sci. Res.* **2013**, *5*, 125–154.
25. Niemeyer, P.; Knudsen, J. *Learning Java*; O'Reilly & Associates: Sebastopol, CA, USA, 2000.
26. Batyuk, A.; Voityshyn, V.; Verhun, V. Software Architecture Design of the Real-Time Processes Monitoring Platform. In: Proceeding of the Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 98–101.
27. Molnár, E.; Molnár, R.; Kryvinska, N.; Greguš M. Web Intelligence in practice. *J. Serv. Sci. Res.* **2014**, *6*, 149–172.