A Change in Linguistic features of a text under Mode shift[[1]](#footnote-1)

Litvinova T.A. ([centr\_rus\_yaz@mail.ru](mailto:centr_rus_yaz@mail.ru))

Voronezh State Pedagogical University, Voronezh, Russia

A problem of authorship attribution, i.e. determining the author of an anonymous text using a set of authors, has received a great deal of attention from researchers all over the world due to the ever growing massive of harmful anonymous texts coming from the Internet. Studies in authorship attribution rely on the assumptions of the stability of idiolectal features (low intraidiolectal variation), although it is well-known that many factors cause~~s~~ intraidiolectal variation (mode, topic, genre, etc.). Another assumption underpinning authorship attribution studies is the ability of idiolectal features to distinguish between authors (i.e. a high interidiolectal variation). Numerous computer papers on authorship attribution show that in a cross-domain (i.e. cross-topic, cross-genre and especially cross-modal) scenario it is extremely difficult to determine the author of a disputed text. Besides, little attention is being paid by computer scientists to the analysis of language features and particularly to the level of their variation. This study aims at analyzing written and spoken (in the form of transcripts) texts by the same authors from a specially designed corpus which provides controlling for other factors of variation to answer the following questions. First, could we automatically discriminate between written and spoken texts and which features contribute to it most? Second, are there any features immune to mode change which are capable of discriminating between authors? We concluded that large-scale corpora representative of an idiolect of a typical Russian speaker as completely as possible are needed to develop more feasible authorship attribution techniques.

**Key words:** idiolect, idiolectal variation, written and spoken mode, authorship attribution, stylometry, cluster analysis.

Изменение лингвистических характеристик текста при смене модуса

Литвинова Т.А. ([centr\_rus\_yaz@mail.ru](mailto:centr_rus_yaz@mail.ru))

Воронежский государственный педагогический университет, Воронеж, Россия

**1. Introduction**

There has been a growing interest towards idiolect as an individual manifestation of a language system among researchers as well as forensic experts. As the use of the Internet as a major communication tool has exploded, there are currently more and more potentially harmful texts posing threats of all sorts including those of extremism, cyberbulling, etc. to be investigated in terms of their authorship. As was rightly pointed out by forensic linguists [7], it is impossible to design evidence-based and viable authorship attribution (AA) methods unless theoretical issues underlying idiolect are looked into. Therefore studies of a combination of factors contributing to interdialectal and intradialectal variation are currently gaining momentum.

One of the recent trends in authorship attribution is cross-domain AA [19; 20; 21], which was the subject of the latest PAN competition - a world-wide known series of scientific events and shared tasks on digital text forensics and stylometry [11; 17]). Research interest in cross-domain AA is caused by a practical demand for methods for determining the authors of anonymous texts when a forensic expert has a few texts with known authorship which differ in domain (genre, topic, style, register, mode etc.) from that being disputed. This task is extremely daunting and this is when idiolect markers with a low intraindividual variation, i.e. immune to changes in domain are essential.

In a paper aimed at comparison of АА model accuracies in different cross-domain scenarios [6] it was noted that AA models perform worst in cross-mode scenario, i.e. when training and test samples include texts of different modes: “This suggests that how people write and how they speak may be somewhat distinct” [6, p. 342]. However, studies in cross-mode АА are very scarce. There are also not enough theoretical studies seeking to discriminate between oral and written texts by the same person when other factors contributing to variation of idiolect are controlled. The objective of the paper is thus to evaluate the effect of the mode (spoken and written) on text features using a corpus of oral and written texts on the same topic collected from the same speakers.

Therefore, we aim to answer two research questions:

1) Which idiolectal features are affected by mode change, i.e. have a high intrainidividual variation in cross-mode scenario?

2) Which text features are immune to mode change and do allow one to distinguish between different idiolects in cross-mode scenario under control of any other factors of variation?

Our contribution to the field is two-fold. First, we broaden the knowledge of the differences between written and spoken discourses. Second, we add to the field of cross-domain AA by proposing a direction to the search of features for developing models useful in cross-domain single-topic scenario.

**2. Мode as a factor of idiolectal variation**

All studies in AA are underpinned by the assumption that idiolectal features are stable (“human stylome” hypothesis by Halteren et al. [9], “individual invariant” set forth by Fomenko and Fomenko [16]). However, it is obvious that this has to be more profoundly investigated [14; 16]. Cross-domain AA studies that suggest that models perform worse when the training and test samples differ in topics and genres indicate implicitly that there are changes in the characteristics of idiolect caused by the above factors. However, little attention is paid (particularly, in “computer” papers) to actually analyzing linguistic features contributing to intraidiolectal variation. What makes studies of the effect of various factors on the stability of the idiolectal characteristics even more daunting is that they require special text corpora that would enable the influence of individual factors on intradialectal variation to be investigated, which is particularly the case for the mode factor. We are only aware of one such a corpus in English [5] used for cross-domain AA [6; 21] and gender identification [18] with no attention being paid to the problem of the stability of certain idiolectal features in case of mode change. Although there are some studies aimed at a systematic comparison of spoken and written discourses (see, for example, [2]), but their results should be taken with caution since spoken and written texts used for comparison differ not only in their mode but also in functional style, register, etc. (as in the work by Chafe cited above [2] where scientific prose and informal everyday speech was contrasted).

A corpus of Russian spoken and written texts by the same authors created by Kibrik [12] allows one to analyze the similarity of spoken and written texts by the same authors controlling for other factors of variation (topic, situation, authors). Although this corpus is rather artificial and small, it is nevertheless unique since it allows to control factors of idiolectal variation and reveal affect of mode shift.

Using this corpus, Kibrik has made a few general conclusions on the overall differences between spoken and written language [12]. He states that more complex, lengthy clauses, terms, normative syntactic structures are characteristic of written texts, whereas shorter clauses, colloquial words and syntactic structures are characteristic of spoken texts, but this observation was made with no statistical data analysis.

Litvinova et al. 2018 [14] used this corpus as well to conclude that mode change causes the largest number of statistically significant differences between written and spoken texts by the same authors in comparison with topic change, text type (story or retelling), time of text production, although there are some features resistant to mode change (most of them are conjunctions). They use the Wilcoxon signed-rank test to identify the stability of features in spoken and written texts by the same author and rejected the null hypothesis about the absence of effect of mode change for such features as mean sentence length in words; percentage of words longer than 6 letters; percentage of negations; percentage of adverb-like pronouns; negation *не* (‘not’); percentage of deictic words; percentage of 100 most frequent words in Russian, as well as the number of features reflecting the frequency structure of the text measured by means of QUITA analyzer [13]. QUITA analyzer allows users to calculate different indices related to vocabulary richness and text complexity both thoroughly studied but length-dependent (for example, TTR) and new indices that are computationally tractable and sample-size invariant. These features are thus affected by mode change.

In this paper we aim to look deeper into the influence of mode on linguistic characteristics of texts and use the discriminant analysis to assess the contribution of individual features to the discrimination between written and oral texts. Next, we analyze the features which are immune to mode change and estimate their ability to discriminate authors in cross-mode single-topic scenario using cluster analysis.

**3. Experimental study of linguistic feature behavior under mode shift**

**3.1. Corpus**

We used a freely available corpus of Russian-language texts “Funny stories” (http://www.spokencorpora.ru/showcorpus.py?dir=02funny). The corpus contains 40 pairs of stories by adults (aged from 18 to 60) telling about funny accidents in their lives. Each participant contributed 2 stories on the same event, i.e. in writing and orally. The data was collected in two stages: first, each participant was instructed to tell their story orally without being told that they are going to be asked to do the same in writing. A week later they were instructed to perform the task in writing. The total length of the oral part of the corpus is about 1 hour 10 minutes with the total of about 7 thousand words (we analyzed provided transcripts). The total length of the written part is about 10 thousand words. Total size of the corpus is 17078 tokens (or 4521 unique word forms).

**3.2. Which texts features are affected by mode change?**

For the first experiment we employed the set of parameters from [14] and carried out the discriminant analysis. The aim of the analysis is to group a variety of objects (texts in this case) into smaller classes based on the discriminant variables (text features). As the number and composition of the classes is originally specified in the discriminant analysis, the main goal is to determine how accurately one can class an object (oral and written texts) using the set of discriminant variables (idiolectal features).

We performed a two-group discriminant analysis in IBM® SPSS® Statistics 25.0. Our research question is: which text features allow us to discriminate between spoken and written texts and which features contribute to the discrimination most?

At the first stage, we applied tests of equality of group means (one-way ANOVA) to measure each independent variable's potential before the model is created. We excluded the variables with the significance >0,05 from the next analysis. Table 1 shows the parameters used for the construction of the model, ANOVA statistics and mean values of variables in written and spoken texts (the smaller the Wilks's lambda, the more important the independent variable is to the discriminant function; parameters are ranked in accordance with the Wilks' Lambda values).

Table 1

Variables used in the discriminant model and their mean values

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Wilks' Lambda | F | Sig. | Mean values in written texts | Mean values in spoken texts |
| Percentage of deictic words | 0,708 | 32,105 | <0,001 | 4,286 | **8,205** |
| Text complexity index Lambda | 0,806 | 18,775 | <0,001 | **1,624** | 1,45 |
| Percentage of words longer than 6 letters (Sixltr) | 0,819 | 17,291 | <0,001 | **29,43** | 24,424 |
| Average token (word) length in characters (Av\_Tok\_Len) | 0,880 | 10,612 | 0,002 | **4,945** | 4,678 |
| Text complexity index Cur\_Len\_RIndex | 0,914 | 7,381 | 0,008 | **0,943** | 0,932 |
| Percentage of 100 most frequent words | 0,918 | 6,954 | 0,010 | 38,107 | **41,659** |
| Text complexity index R1 | 0,927 | 6,171 | 0,015 | **0,911** | 0,889 |
| Text complexity index RRmc | 0,934 | 5,509 | 0,021 | **0,975** | 0,969 |
| Adverb-like pronoun | 0,934 | 5,493 | 0,022 | 2,987 | **4,015** |
| Не “Not” | 0,936 | 5,332 | 0,024 | **1,864** | 1,229 |
| Negations | 0,955 | 3,653 | 0,049 | **2,209** | 1,623 |

The resulting discriminant function has a high eigenvalue ~1,914 and a canonical correlation 0,810 (equivalent to the Pearson's correlation between the discriminant scores and the groups). We used Wilks' lambda as a measure of how well the function separates cases into groups. Its value (0,343, <0.001) is indicative of a good discriminatory ability of the function. The accuracy of classification is 93,8%.

Let us look at the structure matrix which shows the correlation of each predictor variable with the discriminant function. This is another way to assess the contribution of the individual predictors (the higher is the rank, the more significant the predictor is) (Table 2). As suggested elsewhere, the structure matrix is unaffected by collinearity, which is very important for linguistic data.

Table 2

The structure matrix

|  |  |
| --- | --- |
|  | Function |
| 1 |
| Percentage of deictic words | -0,679 |
| Lambda | 0,519 |
| Sixltr | 0,498 |
| Av\_Tok\_Len | 0,390 |
| Cur\_Len\_RIndex | 0,326 |
| 100 most frequent words | -0,316 |
| R1 | 0,298 |
| RRmc | 0,281 |
| Adverb-like pronoun | -0,281 |
| Не “not” | 0,277 |
| Negations | 0,228 |

As we can see, the features are ranked in the structure matrix in the same manner as in Table 1. Therefore, mode change causes changes in linguistic features which can be divided into two groups: the features related to the percentage of the most frequent words of the Russian language in text (1) and those related to text complexity and vocabulary richness (2). As we can see, spoken texts contain more most frequent words in total and deictic words in particular (*там* “there”, *здесь* “here”, *этот* “this”, *те* “those” etc.), although written texts contain more negations. Written texts by the same authors have higher indices of text complexity. They contain more long words (longer than 6 letters), higher average word length index and higher indices of vocabulary richness [13]:

**R1,** an index of vocabulary richness which is based on the Hirsch h-point (and the cumulative relative frequencies up to that h-point);

**RRmc** which is a normalized index of repeat rate, a degree of vocabulary concentration in a text [13].

**Cur\_Len\_R Index**, a vocabulary richness index derived from the curve length (L) (a vocabulary richness index based on the curve of the rank–frequency distribution). It is determined by the ratio of the curve length above the Hirsch h-point to the whole curve length;

**Lambda**, a further development of curve length L to normalize for text length.

Mostly, our results related to the features which are affected by mode change are in good agreement with those obtained in [14]. Using different statistical method, we get the same list of features which change as does the mode. So, it is safely to say that there are intraindividual idiolectal variations caused by mode change. Moreover, we were able to build a classifier which distinguishes between spoken and written texts by the same authors with a very high degree of accuracy using a relatively small set of features. Now we move toward the next research question related to the interindividual variation.

**3.3. Which features immune to mode change allow us to distinguish between different idiolects?**

To assess the level of interindividual variation of the features immune to mode change, we designed boxplots for written and spoken texts separately (the examples on Fig. 1 for “percentage of conjunctions”, “percentage of personal pronouns”).

|  |  |
| --- | --- |
|  |  |

Fig. 1. Examples of boxplots used for evaluating a range of interindividual variation (w – written texts, s – spoken texts)

A visual inspection of the data showed that some of the parameters immune to mode change do range widely in their variation from individual to individual. It should be noted that while some of the features have quite a similar range of variation in two modes (i.e. personal pronouns, as well as noun-like pronoun, *что* “that”), the others have a wider range of variation in one of the modes (the total of the conjunctions in the spoken mode, conjunction *и* “and” and the total of the prepositions in the written mode).

To assess their ability to distinguish between idiolects, we performed a hierarchical cluster analysis (HCA), an exploratory statistical technique that identifies group samples by similarity [4], and dendrograms are used to visualize the results. HCA typically used in corpus linguistics for non-diachronic data [8]. In HCA, “we take the individual data points and in a step-by-step (hierarchical) procedure join (i.e. agglomerate) the closest ones until we create one large cluster containing all the data points” [1, p. 323]. To reveal natural structure in our data, we did not specify the number of clusters prior to the analysis. We assumed, however, that documents by the same author (written and spoken texts) will be combined in one cluster. However, using different clustering methods and combinations of features immune to mode change, we were not able to obtain more than 15% of text pairs clustering by authorship. We can assume that some other factors of “stable” interidiolectal variation (e.g., gender, personality traits) affect at least some of these features but not the authorship.

Searching for idiolectal features which could cluster written and spoken texts by the same authors, we have tried different sets of features – most frequent words (MFWs) of the corpus since they perform well in different text classification tasks as well as most frequent word bigrams and trigrams and character n-grams (with n=2, 3, 4, 5) as features [15]. We used stylo package (version 0.6.9) for feature extraction and clustering in the statistical programming environment R [3]. As in [15], we did not lemmatize corpus for our experiments in order to avoid data sparseness. We also used All features were normalized using z-scores.

We have performed experiments with different numbers and ranges of MFWs, word and character n-grams starting from 10 to 1000. We used Ward linkage as the clustering method since it minimizes the total variance within-a cluster [4]. We also used bootstrap consensus trees [3] as a way to improve the reliability of the cluster analysis dendrograms.

The best results in terms of clustering by authorship were obtained using MFWs as features and Cosine Delta Distance. This is a function for computing a cosine similarity of a scaled (z-scored) matrix of values, e.g. a table of word frequencies. Recent findings [10] show that this distance outperforms other nearest neighbor approaches in the domain of authorship attribution.

It is of interest (see Fig. 2) that with a relatively small number of MFWs (100; most of them are function words) a reasonable number of texts (25 pairs from 40) are clustered by authorship (file names contains metadata, for example 20\_S\_М\_34 means 20(author ID)\_S(spoken text)\_M(male)\_24(age)). It is also to be noted that some of the texts are grouped in larger clusters by author gender and(or) age group; it is our future plan to analyze larger clusters and their structure with a detailed analysis of the features.

However, the use of top frequent words only which are basically function words (one of the most popular features in AA) does not provide perfect author clustering. We run experiments with more MFWs (as well as word and character n-grams) using Bootstrap Consensus Tree (BCT). BCT is the analysis which runs lots of cluster analyses and shows a radial graph with clusters emerging from the centre of the screen. Consensus strength of 0.75 was chosen which means that visualized linkages appear in at least 75% of the clusters.

The best results in terms of author clustering were obtained using 500–1000 MFWs as features (see Fig. 3).



Fig. 2. Cluster analysis with 100 MFWs

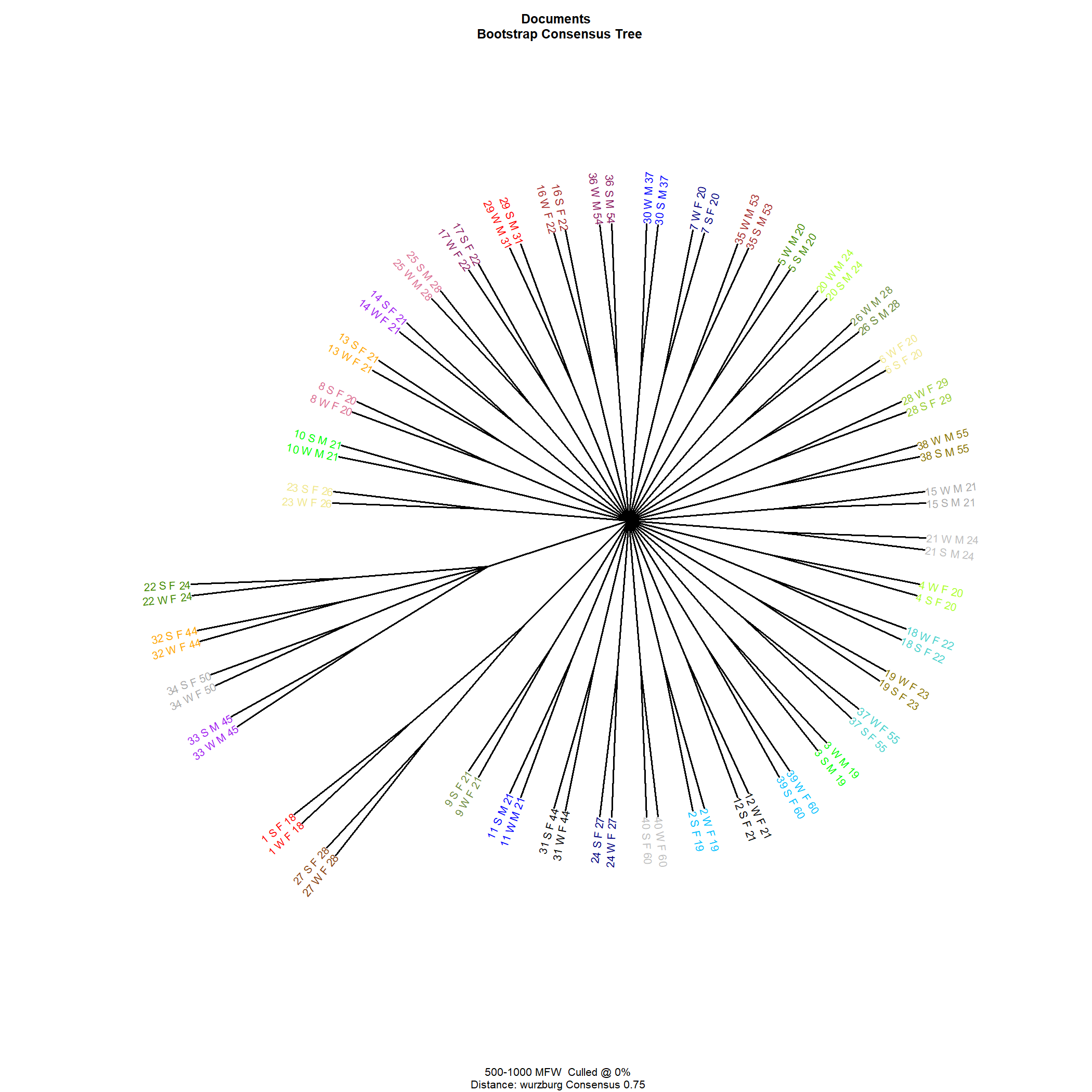


Fig. 3. Bootstrap Consensus Tree with Cosine Delta and 500–1000 MFWs

Even though the texts differ in mode and were produced in different periods of time, the authors remain consistent (to some extent) in their word choice when talking about the same topic. Moreover, texts by different authors could be separated from each other based on MFWs frequencies. Contrary to our expectations, character and word n-grams did not show stability when person telling the same story in different modes.

Our results are in line with findings by Stamatatos [21] who found that high values of MWFs in English provided best results in cross-genre single-topic scenario (both spoken and written texts were analyzed): “Given that all texts are on the same topic, topic specific words provide useful clues of authorship” [21, p. 12]. Thus, it could be the case that universal set of features for AA does not exist, and in different AA scenarios different type of idiolectal features should be useful.

**4. Conclusions and Future work**

Our study shows that written and spoken texts by the same authors on the same topic have systematic statistically meaningful differences which relate mostly to the text complexity. Although there are idiolectal topic-independent features immune to mode change (mostly certain part-of-speech and individual function words), using only a small set of these features did not allow us to discriminate between authors in cross-mode single-topic scenario.

Using most frequent words of the corpus did allow us to separate texts by the same author, with some reasonable results starting from 100 MFWs to perfect separation in the range of 500-1000 (self-evident, this list contains many topic-related words, which turned out to be good authorship markers, - for long time AA studies concentrated mostly on function words). Contrary to our expectations, word and character n-gram features were not useful for author clustering. Only most frequent words of the corpus related to certain story each person were telling were useful in our experiment.

This study has some obvious limitations due to the extremely small corpus size and rather artificial nature. In practice, situation is typical where texts from different modes, genres and topic should be attributed according to their authorship. As we can see, mode shift causes dramatical changes in linguistic features of the text; features immune to the shift do not distinguish authors. Only topic-related words distinguish authors but using these features is not feasible in real-life situation when there is no control for other factors of variation (register, genre, topic, etc.). Therefore, there is a strong need for the search of idiolectal features with high interindividual and low intraindividual variation.

To facilitate this search, we launched the collection of RusIdiostyle Corpus, a special text corpus that represents an idiolect of each speaker as completely as possible and contains texts by the same authors on multiple topics, genres, modes, etc. We argue that corpus-based study of idiolectal variation will not only contribute to our understanding of the phenomenon of individual realization of a language system, but will also help more meaningful and applicable techniques of authorship attribution to be developed.

References

1. Brezina V. (2018), Statistics in Corpus Linguistics: A Practical Guide, Cambridge, Cambridge University Press.
2. Chafe W. (1982), Integration and Involvement in Speaking, Writing, and Oral Literature. In: Tannen D. Spoken and Written Language: Exploring Orality and Literacy, Norwood, Ablex, pp. 35-53.
3. Eder M., Rybicki J., Kestemont M. (2016), Stylometry with R: a package for computational text analysis, R Journal, 8(1), pp. 107-121.
4. Everitt B. S., Landau S., Leese M., Stahl D. (2011), Cluster Analysis, John Wiley & Sons Ltd., Chichester.
5. Goldstein-Stewart J., Goodwin K.A., Sabin R.E., Winder R.K. (2008), Creating and Using a Correlated Corpus to Glean Communicative Commonalities. Proceedings of LREC 2008, available at: <http://www.lrec-conf.org/proceedings/lrec2008/index.html>
6. Goldstein-Stewart J., Winder R., Sabin R. (2009), Person identification from text and speech genre samples, Proceedings of the 12th Conference of the European Chapter of the ACL, EACL, Athens, Greece, pp. 336–344
7. Grant T., Macleod N. (2018), Resources and constraints in linguistic identity performance: a theory of authorship, Language and Law/Linguagem e Direito, vol. 5, pp. 80-96.
8. Gries S. Th. (2013), Statistics for linguistics with R: a practical introduction, Berlin, Walter de Gruyter.
9. Halteren H. van, Baayen R.H., Tweedie F., Haverkort M., Neijt A. (2005), New Machine Learning Methods Demonstrate the Existence of a Human Stylome, Journal of Quantitative Linguistics, vol. 12, pp. 65-77.
10. Jannidis F., Pielstrom S., Schoch Ch., Vitt Th. (2015). Improving Burrows' Delta: An empirical evaluation of text distance measures, Book of Abstracts of the Digital Humanities Conference 2015, ADHO, UWS, Sydney, Australia, available at: <http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html>
11. Kestemont M., Tschuggnall M., Stamatatos E., Daelemans W., Specht G., Stein B., Potthast M. (2018), Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection, Working Notes Papers of the CLEF 2018 Evaluation Labs, Avignon, France, pp. 1-25.
12. Kibrik А. А. (2009), Mode, genre and other parameters for discourse classification [Modus, zhanr i drugie parametry klassifikacii diskursov], [Voprosy jazykoznanija], Vol. 2, pp. 3–21 [in Russian].
13. Kubát M., Matlach V., Čech R. (2014), QUITA: Quantitative Index Text Analyzer”, Studies in quantitative linguistics, vol. 18. RAM-Verlag.
14. Litvinova T., Litvinova O., Seredin P. (2018), Assessing the Level of Stability of Idiolectal Features across Modes, Topics and Time of Text Production. Litvinova et al. 2018, Proceedings of 23rd Conference of Open Innovations Association FRUCT 2018, Bologna, Italy, pp. 223-230.
15. Mandravickaitė J., Krilavičius T. (2017), Stylometric analysis of parliamentary speeches: gender dimension, Proceedings of BSNLP 2017: the 6th workshop on Balto-Slavic Natural Language Processing (EACL 2017), Valencia, Spain, 2017, p. 102-107.
16. Miheev M., Ehrlich L. (2017), Connectors frequencies as a distinctive sign of the individual style (in view of the couple Fomenko hypothesis) [Chastota sluzhebnyh slov kak razlichitel'nyj priznak idiostilya (v svyazi s gipotezoj suprugov Fomenko)], “Computational Linguistics and Intellectual Technologies” [Komp'juternaja lingvistika i intellektual'nye tehnologii], available at: <http://www.dialog-21.ru/media/3976/miheevmehrlichl.pdf> [in Russian].
17. PAN @ CLEF 2019. Cross-domain Authorship Attribution, available at: <https://pan.webis.de/clef19/pan19-web/author-identification.html>
18. Sabin R. E., Goodwin K. A., Goldstein-Stewart J., Pereira J. A. (2008), Gender Differences across Correlated Corpora: Preliminary Results, Proceedings of the Twenty-First International FLAIRS Conference, Miami, Florida, pp. 207-212.
19. Sapkota U., Solorio T., Montes M., Bethard S., Rosso P. (2014), Crosstopic authorship attribution: Will out-of-topic data help?, Proceedings of the 25th International Conference on Computational Linguistics (CoLing): Technical Papers, Dublin, Ireland, pp. 1228–1237.
20. Stamatatos E. (2013), On the robustness of authorship attribution based on character n-gram features, Journal of Law and Policy, vol. 21, pp. 421-439.
21. Stamatatos E. (2017), Masking topic-related information to enhance authorship attribution, Journal of the Association for Information Science and Technology, 69(3), pp. 461–473.

1. The work is supported by the grant of Russian Science Foundation No 18-78-10081 “Modelling of the idiolect of a modern Russian speaker in the context of the problem of authorship attribution”. I would like to thank three anonymous reviewers for their insightful comments. [↑](#footnote-ref-1)