Assessing the Level of Stability of Idiolectal Features across Modes, Topics and Time of Text Production

Tatiana Litvinova, Olga Litvinova
Voronezh State Pedagogical University
Voronezh, Russia
centr_rus_yaz@mail.ru

Pavel Seredin
Voronezh State University
Voronezh, Russia
paul@phys.vsu.ru

***Abstract*—Authorship attribution, i.e. task of revealing the author of a disputed text, is one of challenging issues facing digital forensics. Cross-domain authorship attribution when training and test texts differ in genres, topics and even modes (written/oral) is the most realistic, yet the most difficult scenario. All authorship attribution studies rely on the notion of an idiolect, which is a set of stable features, despite the fact that there are few studies exploring the stability of idiolectal features. The aim of the paper is to reveal the effect of mode, topics and time of text production on the stability of idiolectal features across a series of experiments. Our pilot study revealed that a mode change (written/oral) causes the most striking differences in text parameters in comparison to a topic and time of production although some features (namely, relative frequencies of certain discourse markers) remain relatively stable in all experimental setups. We conclude that the corpus containing diverse types of texts from each individual is needed for thoroughly examining the stability of idiolectal features and developing cross-domain attribution techniques to be employed in realistic scenarios.**

I. INTRODUCTION

In this day due to a rapid growth of Internet communications, text analysis for author identification purposes is gaining particular importance. Long literary texts have mainly been used in this kind of studies while there is currently a growing need to identify the author of Internet texts containing threats, extremist and terrorist propaganda, etc.

However, there are a lot of issues pertaining to identification of individual language styles that remain to be addressed. Methods of authorship attribution based on analyzing literary texts are not suitable for Internet texts as they are not sufficiently long and challenging for natural language processing, etc. [1],[2]. In addition, expert linguists tend to deal with actual threatening letters, social media posts, etc., i.e. texts of known authorship (training texts) and the texts under investigation belong to a different thematic area and (or) different genres. This task is called cross-domain attribution. In recent years, scholars have also started to tackle the problem of such a kind of attribution [3]. This year cross-domain attribution shared task was introduced at PAN, evaluation lab on digital text forensics (https://pan.webis.de/clef18/pan18-web/author-identification.html).

This task has been shown to be a very difficult one [3].

Cross-lingual authorship attribution aimed at "finding the author of an anonymous document written in one language by using labeled documents written in other languages" [4] has also been given a fair amount of attention and there have been attempts to show that it was actually possible to perform [4],[5].

It should be noted that in spite of its obvious theoretical and practical importance, very little research effort has been dedicated to the problem of cross-modal attribution, i.e. that of determining the author of a written text with oral speech samples as training material. In [6] different cross-domain attribution scenarios were analyzed and compared. It was shown that a training/test from different modes (oral/written) was the most difficult scenario: "A key factor in determining which sets can successfully be used to train other sets seems to be the **mode**, that is, whether or not a set is textual or spoken, as the lowest accuracies tend to be found between genres of different modes. This suggests that how people write and how they speak may be somewhat distinct" [6].

It is obvious that for effective cross-modal attribution one should use stylometric features that remain relatively stable over mode change. We are not aware of the studies aimed at the distinction of the most and less stable idiolectual features involving oral and written samples.

This resulted in the aim of the paper, which is to assess the level of stability of idiolectal features across modes (written/oral), as well as topics and time of text production using appropriate text corpora.

II. IDIOLECTAL FEATURES ACROSS MODES, TOPICS AND GENRES

The very idea of authorship attribution is based on the assumption about the existence of unique, distinct versions of a language in a writer (idiolect): "every native speaker has their own distinct and individual version of the language they speak and write" [7]. This "individual version" supposed to be consistent across texts of different topics, genre etc., but this remains speculative in the absence of systematic empirical investigation [8]. As Grant and Macleod claim, "in fact no

studies seem to exist that would demonstrate individual consistency 'across genre and modes of production" [9].

Studies dealing with cross-domain and even cross-lingual attribution are an indirect indication that there is some consistency (which enables one to argue there is "a human stylome", see, e.g., a paper [10]), although "whether the differences between individual language forms are equally visible for all aspects of a language" [10] is yet to be addressed. Special investigations are needed aimed at assessing the level of stability of idiolectal features depending on certain characteristics (genre, topic, mode, social context, etc.) since "persistence of identity therefore does not require a static and unchanging identity. It does, however, require more understanding about which aspects of identity performance remain stable while the resources we draw on are changing in each specific interactional moment" [9].

One of the fundamental characteristics of text production is its mode. Studies of fundamental differences between oral and written texts have long been carried out by linguists, however few contained comparisons of oral and written texts by the same individual rather than a massive of oral and written texts by different people. DeVito [11] compared written and spoken speech samples of ten professors on topics of professional interest. He found their writing contained longer and less common words, as well as a larger diversity of words. Driemann [12] obtained similar results while analyzing written and spoken descriptions of paintings by graduate students. More recent study [13] looking at morphosyntactic (number of subordinate relative pronouns, the use of modal verbs, etc.) and discursive-pragmatic variables of speaker's personal style in oral and written discourse of the same individuals has shown that units of the former level are more consistent in the idiolect of an author than of the latter one.

Cross-topic authorship attribution studies widely use function words (i.e., prepositions, articles, etc.) as features [14]. However, Mikros and Argiri [15] demonstrated that these features are not immune to topic shifts.

Other types of features found to be effective in cross-topic and cross-genre AA are Linguistic Inquiry and Word Count (LIWC) features [6], although in the study [16] aimed at assessing the level of stability of LIWC features in different texts by the same authors, it was shown that most of LIWC features demonstrate high intra-variability. It should be noted that using LIWC features yields low accuracy for the task "Predict Communicant in a Speech Genre Given Information on Textual Genres", i.e. in cross-modal scenario [6] which means that mode and topic could effect in different ways on the same linguistic features.

Character n-grams [17],[18] are widely used in cross-domain attribution but suffer from lack of linguistic theory behind them. N-grams containing punctuation marks show the best performance in the cross-topic scenario in [17]. Punctuation mark frequencies are used as well [19].

Features reflecting the lexical complexity of a text were used in the study aimed at finding cross-lingual correlates of idiostyle [20] with some strong correlations revealed.

It should be thus acknowledged that there has been no agreement to date as to which parameters of idiolect are stable. We argue, however, that such parameters must be highly frequent and easily extracted. For attribution tasks it is also important that a parameter was capable of distinguishing texts by different authors as well as combining texts by the same author, i.e. showing high inter-speaker and low intra-speaker variability.

Despite the obvious spark of research interest towards idiolect as a result of the manifestation of the properties of a language system in individual speech, there are still a lot of issues that need to be addressed. There has been no comprehensive approach to such important problems as a correlation of the mechanisms of changeability and stability of style, comparison of the character of variation of properties of different linguistic levels and aspects, limitations of dynamic variation of the parameters and their scope as well as interaction of linguistic characteristics.

We are contributing to the field by analyzing the level of stability of a range of highly frequent idiolectal features in different types of writing and speech samples of the same speakers. To be more specific, we are exploring the effect of change of mode, topic and time of production on idiolectal features (separately) while controlling the other variables. The level of inter-speaker and low intra-speaker variability parameters has also been evaluated.

III. Experimental setup

A. Mode change

The first experiment is aimed at assessing the level of stability of idiolectal features across modes of text production. 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. The total length of the written part is about 10 thousand words. The average length of an oral text is 251 words (SD=221), and of a written one 177 words (SD=120).

These documents were unedited, actual writings; they contain different numbers of words and are rather short. These experiments are intended to test the lower limits of text length and quantity "because forensically significant documents are often short and cannot be amplified; indeed, even known documents are often short in length and limited in quantity" [21]. Thus, we are interested in analyzing short texts, since "in fact, it is important to develop techniques which can operate successfully on short documents, as the worst case scenario" [21].

Demographics of the authors of the texts from "Funny stories" corpus are presented in Table I.

TABLE I. DEMOGRAPHICS OF THE AUTHORS OF FUNNY STORIES CORPUS

	Gender	
Age	Female	Male
18-21	9	6
22-29	9	4
31-44	2	2
45-54	1	3
55-60	3	1

The corpus contains oral narratives that are available for listening as well as their transcripts (minimum and abridged versions). In the current study we used the abridged version of the labeling where sentences as well as elementary discourse units (EDU) are marked. From the physiological point of view, EDU is pronounced in one breath. From the cognitive point of view, one "focus of consciousness" is verbalized, i.e. all the information that selective human consciousness is capable of retaining. Linguistically speaking, semantic capacity of a canonic EDU is a description of an event or state. From the syntactic viewpoint, such a canonic EDU is a predicate (clause). Finally, the prosody of EDU is a phonetic contour in terms of the tone movement (frequency), main accent center (intensity), tempo (acceleration - delay) and volume (attenuation by the end) [22].

Any comments that were not made by the main storyteller as well as interruptions were removed. Words with non-standard phoneme realisations were transformed into a standard form.

We used several groups of features:

1) ***General linguistic features***: mean sentence length in words (**WPS**); percent of words longer than 6 letters (**Sixltr**) of the total number of words; percent of periods (**Period**).

2) **POS features:** percent of adverbs (ADV), prepositions (PREP), conjunctions (CONJ), negations (NEG) as well as different types of pronouns: noun-like (*кто* "who", *я* "I"), adverb-like (*когда* "when", *где* "where"), adjective-like (*мой* "my", *наш* "our"), personal (*я* "me").

3) **FW_{separ}:** percentage of most frequent Russian words in accordance with [23]: И "and", В "in", НЕ "not/no", НА "on", С "with", ЧТО "what", А "but".

4) **DEIC:** percentage of deictic words (*там* "there", *здесь* "here", *этот* "this", *те* "those" etc.).

5) **DM:** percentage of discourse markers. These words are referred to more commonly as 'linking words' and 'linking phrases', or 'sentence connectors'. Discourse markers structure speech as well as mental processes that are controlled by the speaker. We have designed a dictionary of discourse markers which consists of 13 semantic groups of markers with one marker falling into more than one category. However, in this study we employ only three groups of the most frequent DMs: conjunction (*и, тоже, даже*), addition (*вдобавок, после этого, вскоре*), contrast (*но, хотя, несмотря, вопреки*).

6) **FREQ.** The percentage of 100 most Russian frequent words. The list of these words was taken from [23]. Most of them are function words.

7) **QUITA:** indicators of lexical complexity and frequency structure of a text. We calculated them using Quantitative Index Text Analyser (QUITA) tool [24]. Detailed examples of computations of most indicators used in QUITA can be found in [25]. Although this software is new, it has already been used in several studies dealing with quantitative analysis of genres [26], stylometric analysis of inaugurational speeches [27], cross-linguistic authorship attribution [20], etc. Tokenization was performed using a built-in tokenizer; no lemmatization has been performed.

From a set of variables which can be calculated using QUITA, we selected only those which are relatively text-independent [24]:

- **Average Tokens Length** is the arithmetic mean of the lengths of tokens. As Juola et al. [20] state, this is one of the earliest methods of assessing lexical complexity.
- **R1** is an indicator of vocabulary richness which is based on the h-point (fuzzy boundary point on the curve where the rank is the same as the frequency), but it reduces the impact of text length.
- **RRmc** is the relative repeat rate (RR) which puts the results in the interval <0;1>. RR shows the degree of vocabulary concentration in a text.
- **Λ** is an indicator which deals with a frequency structure of text. "On the one hand, the lambda is related to vocabulary richness, and on the other hand, it takes into account the relationship between neighbouring frequencies" [24].
- **WritersView** is an indicator connected to the golden ratio. "It is supposed that each author of any text must abide by some universal law, namely the golden ratio. The writer's view is defined by the angle between the h-point and the ends of the rank-frequency distribution. The results should approximate to the value of the golden ratio" [24]. Popescu et al. [28] claim that it was 'baptized in this way because one can imagine the writer "sitting" at this point and controlling the equilibrium between autosemantics and synsemantics'.
- **CurveLength R Index** is the ratio of the curve length above the h-point to the whole curve length (curve of rank-frequency distribution).

The stability of a parameter in two samples is normally determined by means of correlations between two pairs of samples of a parameter [13]. However, this approach might be erroneous for the following reasons. It is well known that in order for a correlation to be identified, there have to be a linear dependence between two values. In this case they are values of the same parameter in two texts of the authors. If there is no such a dependence or it takes another form rather than a linear one, it will not be possible to identify a significant correlation. A large dispersion (variation) of a parameter also has a negative effect on determining a correlation.

In this study we employed another approach to determining the stability of the parameters of texts written by a group of

authors which is similar to the approach proposed by Chaski [21] who used chi-square technique to assess the similarity of text samples. As both samples of the same parameter are paired (a property is measured for the same participants), we used Wilcoxon signed-rank test to identify the stability of a parameter in two texts by the same author. This is a non-parametric alternative to paired t-test used to compare paired data. We formulated our hypothesis as follows:

H_0: There is no effect of mode change ($Var_{oral} = Var_{wr}$).

H_1: There is effect of mode change (Var_{oral} not equal to Var_{wr}).

Note that due to a small size of the sample, small frequencies are expected while calculating z-statistics of the parameters. The critical level *p* will be estimated to be extremely low. In order to make up for "optimism" of the criterion *z,* Yates' continuity correction was introduced to minimize the level of error while calculating the critical value of p. In this study, we accept the significance level of $p < 0.05$. We accept this null hypothesis if the probability associated with the test result is greater than 0.05. We reject the null hypothesis accepting instead the alternative hypothesis of difference (effect of mode), if the probability is less than or equal to 0.05. A low p-value demonstrates a significant difference, while a high p-value indicates similarity (stability) or at least consistency [21].

B. Topic change

In the second experiment we attempted to evaluate the effect of a topic on idiolectal parameters.

H_0: There is no effect of topic change.

H_1: There is effect of topic change.

We used a freely available corpus of oral texts collected from native speakers of Russian as part of a two-stage experiment (http://www.spokencorpora.ru/showcorpus.py?dir=03pands/rus). At the first stage the participants were asked to choose one of two pictures "Presents" and "Skiing" to describe. For each set they were given a few seconds to take a look at the pictures and then they were to write a description having the pictures in front of them (text type – narrative). At the second stage 6 or 8 hours following the first one, the participants were instructed to retell the same stories without looking at the pictures (text type – retelling).

The corpus contains 20 stories and 20 retellings by nine native speakers of Russian. The recordings were made in Moscow in 2003–2004. All the participants are Moscow residents — 7 females and 3 males aged from 20 to 30 at the time of recording. The total time of the recordings is about 35 minutes; the corpus has the total of 4.5 thousand words. The mean text length is 108 words (SD=39 words).

We used the same features as well as an additional type of features specific to oral speech. These are features based on EDUs:

1) Mean length of EDUs in words.

2) Number of frequent words divided by the number of EDUs.

3) Number of deictic words divided by the number of EDUs.

4) Number of adverbs divided by the number of EDUs.

5) Number of prepositions divided by the number of EDUs.

6) Number of conjunctions divided by the number of EDUs.

7) Number of negations divided by the number of EDUs.

We used the same methodology as we did in the first experiment. Two series of a pairwise comparison have been made (with controlling for text type – narrative and retelling): 1) "Skiing" Retelling – "Presents" Retelling; 2) "Skiing" Story– "Presents" Story.

C. Time of text production change

In the third experiment we attempted to evaluate the effect of the time of writing on the text parameters (with control for topic). We used the same corpus as we did in the second experiment ("Picture description"). As it was already described, we have two types of texts in this corpus: narratives produced immediately after picture introduction ("narratives") and 6 or 8 hours later ("retellings").

H_0: There is no effect of time of production.

H_1: There is some effect of time of production.

Two series of a pairwise comparison have been made (with controlling for text topic):

1) "Skiing" Retelling – "Ice-Skating" Story.

2) "Presents" Retelling – "Presents" Story.

D. Interindividual and intraindividual variability of idiolectal features

In the last experiment we assessed the level of variability of the parameters in the texts of the corpus (interindividual variability) and in the texts of the same authors (intraindividual variability) using a variation coefficient. The variation coefficient equals a ratio of a standard deviation to the average value. A parameter with a high intervariability and low intervariability is "perfect" for authorship attribution [29].

For the texts of the first corpus only interindividual variability of idiolectal features was evaluated as it only has two texts by the same author.

For the texts of the corpus *Picture description* containing 4 texts by the same author apart from the interindividual variation coefficient, the intraindividual one was also calculated.

The intraindividual variation coefficient is calculated by computing variation coefficients for each author and it is then averaged by the number of authors.

IV. Results and Discussion

Table II contains the results of the analysis aimed at revealing the effect of mode change on idiolectal features.

TABLE II. RESULTS OF THE COMPARATIVE ANALYSIS OF WRITTEN AND ORAL TEXTS OF THE SAME AUTHORS

Parameters	Median (written/oral)	Z	p
1. General			
WPS^{o}	14.29/14.73	1,95	0,05
Sixltr***	29,77/24,580	-4,56	<0,001
Period	7,1/6,68	-1,14	0,26
2. POS			
ADV^{o}	2.63/2.72	1,922	0.05
PREP	14.47/14.34	0,538	0.6
CONJ	9.88/10	0,363	0.72
NEG**	1.99/1.62	-2,637	0.009
pronoun-noun	6.93/7.82	0,941	0.35
pronoun-adverb*	2.92/3.72	2,325	0.02
pronoun-adjective	1.24/1.37	1,124	0.26
personal pronouns	5.17/5.62	1,015	0.31
3. FW$_{separ}$			
И	*4.01/3.9*	*-0,927*	*0,36*
В^{o}	2.83/2.31	-1,898	0.06
НЕ**	1.85/1.08	-2,922	0.004
НА	*1.78/1.86*	*-1,130*	*0,26*
С	*1,23/0.92*	*-1,461*	*0.15*
ЧТО	*1.4/1.69*	*0,595*	*0,56*
А	*1.1/0.8*	*-0,094*	*0.93*
4. DEIC			
DEIC***	4.09/7.48	5,080	<0,001
5. DM			
conjunction	*5.03/5.04*	*0,391*	*0,70*
addition	*4.32/3.97*	*-0,484*	*0.63*
contrast	*1.77/1.94*	*0.3*	*0.77*
6. FREQ			
FREQ**	37,94/43,07	3,132	0,002
7. QUITA			
Average Tokens Length***	4.84/4.72	-4,019	<0,001
R1***	0.91/0.89	-3,212	0,001
RRmc***	0.97/0.96	-3,495	<0,001
Λ***	1.63/1.5	-4,059	<0,001
WritersView	*2.38/2.24*	*0,215*	*0,835*
CurveLength R Index***	0.94/0.93	-3,212	0,001

Note. ***p≤ 0,001; **p (0,001; 0,01]; *p (0,01; 0,05); ^{o}p near significance level [0.05; 0,1]. The parameters for which we can not reject the null hypothesis ("stable") are presented in italics.

The results of the comparative analysis of the texts with different topics are presented in Table III.

TABLE III. RESULTS OF THE COMPARATIVE ANALYSIS OF TEXTS WITH DIFFERENT TOPICS

Parameters	**"Skiing" Retelling – "Presents" Retelling, p**	**"Skiing" Story – "Presents" Story, p**
1. General		
WPS	*0,322*	*0,922*
Sixltr	*0,193*	*0,020**
Period	*0,922*	*1,000*
2. POS		
ADV	*0,275*	*0,193*
PREP	*0,006***	*0,064^{o}*
CONJ	*0,910*	*0,922*
NEG	*0,232*	*0,232*
pronoun-noun	0,049*	0,193
pronoun-adverb	*0,432*	*0,652*
pronoun-adjective	*0,910*	*0,844*
personal pronouns	0,375	0,1^{o}
3. FW$_{separ}$		
И	*0,695*	*0,492*
В	0,004**	0,846
НЕ	*0,275*	*0,232*
НА	0,006**	0,006**
С	0,016*	0,004**
ЧТО	0,027*	0,203
А	*0,844*	*0,156*
4. DEIC		
DEIC	*0.16*	*0.432*
5. DM		
conjunction	*0,375*	*0,131*
addition	*0,770*	*0,492*
contrast	*0,910*	*0,570*
6. FREQ		
FREQ	*0.375*	*0,432*
7. QUITA		
Average Tokens Length	0,014*	0,049*
R1	*0,432*	*0,131*
RRmc	0,193	0,1^{o}
Λ	*1,000*	*0,322*
WritersView	*0,695*	*0,193*
CurveLength R Index	*0,922*	*0,375*
8. EDU		
EDU/WC	*0,375*	*0,375*
Freq_total / EDU	*0,375*	*0,570*
Deictic / EDU	0,1^{o}	0,557
Adverb / EDU	*0,375*	*0,322*
Preposition / EDU	0,064^{o}	0,084^{o}
Conjuction / EDU	*1,000*	*0,625*
Negation/ EDU	*0,160*	*0,426*

Note. ***p≤ 0,001; **p (0,001; 0,01]; *p (0,01; 0,05); ^{o}p near significance level [0.05; 0,1]. The parameters for which we can not reject the null hypothesis ("stable") are presented in italics.

Table IV contains the results of the analysis of oral texts produced at a different time.

TABLE IV. RESULTS OF THE COMPARATIVE ANALYSIS OF TEXTS PRODUCED IN DIFFERENT TIME

Parameters	**"Skiing" Retelling – "Skiing" Narrative, p**	**"Presents" Retelling – "Presents" Narrative, p**
1. General		
WPS	*0,846*	*0,375*
Sixltr	*0,193*	*0,492*
Period	*0,922*	*0,275*
2. POS		
ADV	0,084^{o}	0,922
PREP	*0,625*	*0,770*
CONJ	*0,426*	*0,193*
NEG	*0,301*	*0,275*
pronoun-noun	*1,000*	*0,695*
pronoun-adverb	0,232	0,01**
pronoun-adjective	*0,193*	*0,469*
personal pronouns	*0,922*	*0,322*

3. FW_{separ}		
И	*0,922*	*0,131*
В	*0,250*	*0,232*
НЕ	*0,426*	*0,492*
НА	*0,570*	*0,275*
С	*0,742*	*0,813*
ЧТО	*1,000*	*0,232*
А	*0,297*	*0,875*
4. DEIC		
DEIC	*0.652*	*0.131*
5. DM		
conjunction	*0,770*	*0,922*
addition	*1,000*	*0,322*
contrast	*0,910*	*0,652*
6. FREQ		
FREQ	*0,846*	*1,000*
7. QUITA		
Average Tokens Length	*0,492*	*0,922*
R1	*0,131*	*0,557*
RRmc	0,064º	0,492
Λ	*0,131*	*0,695*
WritersView	*0,432*	*0,432*
CurveLength R Index	0,014*	1,000
8. EDU		
EDU/WC	*0,492*	*0,375*
Freq total / EDU	*0,432*	*0,557*
Deictic/ EDU	0,557	0,037*
Adverb / EDU	0,084º	0,557
Preposition / EDU	*0,375*	*0,695*
Conjuction / EDU	*0,322*	*0,232*
Negation/ EDU	*0,432*	*1*

Note. ***p≤ 0,001; **p (0,001; 0,01]; *p (0,01; 0,05); ºp near significance level [0.05; 0,1]. The parameters for which we can not reject the null hypothesis are presented in italics.

Table V summarizes the results of the analysis aimed at revealing the level of variability of the idiolectal features for both corpora.

TABLE V. VARIABILITY OF THE IDIOLECTAL FEATURES

Parameters	I Exp (written/oral) Inter-individual coefficient of variation	II Exp (oral only)	
		Intra-individual coefficient of variation	Inter-individual coefficient of variation
1. General			
WPS	32/53	9±7	37
Sixltr	17/23	8±4	20
Period	***55/39***	***9±6***	***39***
2. POS			
ADV	62/56	32±19	84
PREP	27/24	15±9	41
CONJ	***28/32***	***15±12***	***43***
NEG	69/77	26±14	64
pronoun-noun	46/44	15±11	44
pronoun-adverb	63/50	29±16	81
pronoun-adjective	100/66	51±33	100
personal pronouns	53/49	15±10	38
3. FW_{separ}			
И	***44/42***	***17±15***	***53***
В	50/67	16±13	50
НЕ	77/82	27±15	58
НА	71/62	26±16	74
С	76/83	30±33	100
ЧТО	89/83	36±43	100
А	89/83	36±43	100
4. DEIC			
DEIC	53/67	20±13	73
5. DM			
conjunction	***35/34***	***12±8***	***43***
addition	***38/36***	***18±14***	***49***
contrast	***63/62***	***34±34***	***70***
6. FREQ			
FREQ	14/16	5±3	13
7. QUITA			
Average Tokens Length	8/7	2±1	6
R1	4/5	1±0.7	3
RRmc	1/1	1±0.1	1
Λ	8/9	2±1	6
WritersView	***15/14***	***5±3***	***13***
CurveLength R Index	2/2	1±0.5	2
8. EDU			
EDU/WC	**-**	***8±7***	***20***
Freq/EDU	**-**	***11±8***	***29***
Deic/EDU	**-**	20±12	73
ADV/EDU	**-**	32±20	84
PREP/EDU	**-**	15±10	41
CONJ/EDU	**-**	***15±12***	***43***
NEG/EDU	**-**	***26±14***	***64***

Note. The parameters which have shown to be "stable" across modalities, topics, and text type are given in bold italics.

The statistical analysis allowed us to conclude that the most parameters for which we reject the null hypothesis belong to the texts with different modes. Plausibly, this can be interpreted that mode causes instability, i.e. the most dramatic differences in idiolectal features. This could be one of the explanations of the results obtained in [6] about the extreme complexity of cross-modal attribution.

We have found out that typically, in written texts in comparison with oral texts of the same authors there are more longer words, more negations, less deictic elements and frequent words which is consistent with some previous findings [30]. We have also revealed that all the parameters of lexical complexity differ significantly in oral and written texts by the same authors except the parameter *Writer's view*.

Time of text production seems to cause the least effect on the analyzed text parameters in comparison with mode and topic.

As our statistical analysis has shown, the share of periods, separate conjunctions, conjunctions overall and discourse markers of different groups are relatively stable across different types of texts of the same authors (low intravariability) while displaying a high interindividual variability. We can claim that the stability relates to making boundaries and choosing means of combining them, which is similar to the findings made by Chaski on a written corpus using syntactically marked punctuation as features [21]. The necessity of analysis of discourse markers in studying an idiolect was emphasized by Kredens who states that "different use of discourse markers is interesting for an additional lesson that it teaches" [31].

As for oral speech, we discovered that a range of linguistic features which reflects the characteristics of elementary discourse units (EDU) are relatively stable across texts, namely mean length of EDU in words, number of frequent words, conjunctions and negations. It is interesting that a similar approach applied by Soler and Wanner [32] to the analysis of literary genre dataset where the frequency of each discourse

relation per EDU (which are understood more broadly, as a minimal unit of any type of discourse) was used as a feature has demonstrated that "discourse features, which have been largely neglected in state-of-the-art proposals so far, play a significant role in the task of gender and author identification and author verification" [32]. We argue that EDU-level indicators should be used more widely in future stylometric research.

Indicators based on lexical complexity should be used for cross-domain authorship attribution with caution since we can see that they are mode-sensitive except *Writer's view* which needs further investigation.

Limitations. The current study is a pilot one and has a range of limitations that are caused firstly by a small number of participants and a small number of texts by them.

Secondly, different demographics of the authors which definitely affect linguistic performance was given no consideration.

Thirdly, inability to reject the null hypothesis is not the same as "proving" this hypothesis (in our case about the stability of linguistic parameters), but we argue that parameters for which high values of probability level have been obtained are definitely worth further consideration.

V. Conclusion and Future Research

We presented an ongoing work on analyzing the level of stability of frequent idiolectal features across modes, topics and time of production. We have revealed that the mode (written/oral) causes the most dramatic differences in text parameters, while texts on the same topic and of the same mode produced at a different time show less diversity. Our analysis revealed that conjunctions and more broadly discourse markers ("linking words", connectors) as well as the share of periods are rather stable across texts by the same authors while showing a high intervariability. For oral texts, the same holds true for a range of markers on the level of elementary discourse units. However, for more substantial conclusions additional research is necessary taking into account the limitations described above.

In our present study we considered mode as a binary category based on a different code (graphic or phonic), however, as new genres appear, the boundaries between written and oral modes become not so clear-cut (i.e. social-media posts). Keeping this in mind, it would be interesting to consider mode as a different position within the continuum of two poles: communicative distance vs immediate communication [33].

Studies on the stability of idiolectal features are limited due to lack of appropriate corpora that should consist of texts on multiple topics, genres, etc. by the same authors. Thus, we are planning to extend this work to larger text collections. For this purpose, we launched the collection of RusIdiostyle Corpus, a special text corpus that represents a discourse space of each participant as completely as possible. It is known for a fact that in order to get insight into the diversity of a discourse, no less than four parameters – mode, genre, functional style and formality – have to be considered [34]. Apart from the above, a parameter showing whether an author has (no) intention to deception. It is of fundamental importance that a corpus has discourses by individuals of different demographics, education levels, personality traits, etc.

Another significant factor is that there are not only texts written according to the researchers' instructions but also "natural" texts (transcripts of recordings of everyday speech, samples of their social media texts, etc.) should be presented in the corpus. Besides, texts should be produced at different time periods.

Another direction of future research involves taking into account the effect of demographics and psychological and cognitive characteristics of the authors on the stability of idiolectal features, which will be possible due to the rich metadata of RusIdiostyle Corpus.

This will help us to design models of an idiolect that would contain stable and variable quantifiable linguistic parameters with various levels of distinctive ability. Based on it, we will also be able to develop an experimental technique for cross-domain authorship attribution to be employed for current and emerging cyber threats.

Acknowledgment

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".

We sincerely thank the three anonymous reviewers for their insightful comments and suggestions.

References

[1] A. Rocha et. al. "Authorship Attribution for Social Media Forensics", in *IEEE Transactions on Information Forensics and Security*, vol. 12, 2016, pp. 5-33.

[2] R. Kaur, S. Singh, H. Kumar, "Authorship Analysis of Online Social Media Content", in *Proc. of 2nd International Conference on Communication, Computing and Networking*, 2019, pp. 141-165.

[3] E. Stamatatos, "Masking topic-related information to enhance authorship attribution", *Journal of the Association for Information Science and Technology,* vol. 69(3), 2017, pp. 461–473.

[4] R. Sarwar, Q. Li, T. Rakthanmanon, S. Nutanong, "A scalable framework for cross-lingual authorship identification", *Information Science,* vol.465, Oct. 2018, pp. 323-339.

[5] D. Bogdanova, A. Lazaridou, "Cross-Language Authorship Attribution", in *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* May 2014, pp. 2015-2020.

[6] J. Goldstein-Stewart, R. Winder, R. Sabin, "Person identification from text and speech genre samples", *in Proc. of the 12th Conference of the European Chapter of the ACL, EACL*, Mar 2009, pp. 336–344.

[7] M. Coulthard, "Author identification, idiolect, and linguistic uniqueness", *Applied Linguistics*, vol. 24, 2004, pp. 431–447.

[8] M. Kestemont, K. Luyckx, W. Daelemans, T. Crombez, (2012). "Cross-Genre Authorship Verification Using Unmasking", *English Studies*, vol. 93, pp. 340–356.

[9] T. Grant, N. Macleod, "Resources and constraints in linguistic identity performance: a theory of authorship", *Language and Law/Linguagem e Direito,* vol. 5, 2018, pp. 80-96.

[10] H. Van Halteren, H. Baayen, F. Tweedie, M. Haverkort, A. Neijt, "New Machine Learning Methods Demonstrate the Existence of a Human Stylome", *Journal of Quantitative Linguistics*, vol. 12, 2005, pp. 65–77.

[11] J. A. DeVito, "Psychogrammatical factors in oral and written discourse by skilled communicators", *Speech Monographs*, vol. 33(1), 1966, pp. 73–76.

[12] G. Drieman, “Differences between written and spoken language: An exploratory study. I. quantitative approach”, Acta Psychologica, vol. 20, 1962, pp. 36–57.
[13] J. A. Garazi, “An analysis of authorship attribution: Identifying linguistic variables in oral and written discourse”, Master thesis. Madrid: Universidad de Madrid, 2016.
[14] M. Kestemont, “Function Words in Authorship Attribution From Black Magic to Theory?”, in *Proc. of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL*, 2014, pp. 59–66.
[15] G. K. Mikros, E. K. Argiri, “Investigating topic influence in authorship attribution”, in *Proc. of the SIGIR'07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, 2007, pp. 29–35.
[16] T. Litvinova, P. Seredin, O. Litvinova, T. Dankova, O. Zagorovskaya, “On the Stability of Some Idiolectal Features”, in Karpov A., Jokisch O., Potapova R. (eds), *Speech and Computer. SPECOM 2018*, Sept 2018, pp. 331-336.
[17] U. Sapkota, T. Solorio, M. Montes, S. Bethard, P. Rosso, “Cross-topic authorship attribution: Will out-of-topic data help?”, in *Proc. of the 25th International Conference on Computational Linguistics (CoLing): Technical Papers,* 2014, pp. 1228–1237.
[18] E. Stamatatos, “On the robustness of authorship attribution based on character n-gram features”, *Journal of Law and Policy*, vol. 21, 2013, pp. 421–439.
[19] H. Baayen, H. van Halteren, A. Neijt, F. Tweedie, “An experiment in authorship attribution”, in *Proc. of 6th JADT*, Mar 2002, pp. 29–37.
[20] P. Juola, G. Mikros, S. Vinsick, “Correlations and Potential Cross-Linguistic Indicators of Writing Style”, *Journal of Quantitative Linguistics*, May 2018, https://www.tandfonline.com/doi/abs/10.1080/09296174.2018.1458395?journalCode=njql20
[21] C. Chaski, “Empirical evaluations of language-based author identification techniques”, *Forensic Linguistics,* vol. 8, 2001, pp. 1-65.
[22] A.A. Kibrik, V.I. Podlesskaja (red.) Rasskazy o snovidenijah: Korpusnoe issledovanie ustnogo russkogo diskursa. Moscow: JASK, 2009 [in Russian].
[23] O. N. Ljashevskaja, S. A. Sharov, *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialah Nacional'nogo korpusa russkogo jazyka),* Moscow: Azbukovnik, 2009 [in Russian].
[24] M. Kubát, V. Matlach, R. Čech, “QUITA: Quantitative Index Text Analyzer”, *Studies in quantitative linguistics,* vol. 18, 2014.
[25] R. Čech, I.-I. Popescu, G. Altmann, *Metody kvantitativní analýzy (nejen) básnických textů*, Univerzita Palackého, 2014.
[26] J. Mandravickaitė, T. Krilavičius, K. L. Man, “Initial Analysis of Characteristics of Textual Genres: Comparison of Lithuanian and English”, in *Proc. of the International MultiConference of Engineers and Computer Scientists,* Mar. 2018, Hong Kong.
[27] M. Kubát, R. Cech, “Quantitative Analysis of US Presidential Inaugural Addresses”, *Glottometrics,* vol. 34, 2016, pp. 14-27.
[28] Popescu, I.-I., et al., *Word frequency studies*. Berlin: Mouton de Gruyter, 2009.
[29] F. J. Nolan, “Auditory and acoustic analysis in speaker recognition”, in Gibbons, J. (ed.), *Language and the law*. London/New York: Longman. 1994, pp. 326-345.
[30] E. van Miltenburg, R. Koolen, E. Krahmer, “Varying image description tasks: spoken versus written descriptions”, in *Proc. of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, Aug 2018, pp. 88–100.
[31] K. Kredens, *Forensic Linguistics and the Status of Linguistic Evidense in the Legal Setting*, PhD dissertation, Lodz, Poland: University of Lodz. 2001.
[32] J. Soler-Company, L. Wanner, “On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification”, in *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* vol. 2, Apr 2017, pp. 681–687.
[33] P. Koch, W. Oesterreicher, “Langage oral et langage ecrit”, in *Lexicon der Romanistischen Linguistik*. Tübingen: Max Niemeyer Verlag, 2001. Tome 1-2, pp. 584-627.
[34] A. A. Kibrik, “Modus, zhanr i drugie parametry klassifikacii diskursov”, *Voprosy jazykoznanija*, vol. 2, 2009, pp. 3–21 (in Russian).