Process-Oriented Characteristics of an Idiolect for Authorship Attribution of Heterogeneous Texts: a Pilot Study

Tatiana Litvinova[1][0000-0002-6019-3700]

[1] RusProfiling Lab, Voronezh State Pedagogical University, 86 Lenina st., Voronezh 394043, Russia
centr_rus_yaz@mail.ru

Abstract. Currently, the task of identification of the author of a typed text is approached mostly in two ways: by means of linguistic analysis (stylometric approach) and analyzing typing behavior (keystroke dynamics approach). The studies which combine these approaches by analyzing complex, process-oriented idiolectal features, although potentially feasible, remain rare. Moreover, existing research focuses mostly on one communication task, thus the question of stability of such features in one's idiolect remain open. The paper presents the results of a pilot study aimed at assessing discriminative ability of two groups of idiolectal markers – non-sequential and sequential process-oriented markers – both separately and in combination in a dataset of heterogeneous texts (dialogues and monologues of different genres) produced by the same writers whose typing process was video-recorded and afterwards manually annotated. The analysis conducted with the use of state-of-the-art methods of identifying the structure in multivariate data and its visualization widely applied in "omics" studies (multilevel PCA, PLS-DA, DIABLO) which have a lot in common with idiolectal studies (a small sample size, a large number of highly correlated predictors) has shown that, despite a strong text-type effect, product-oriented features could discriminate between authors of a short typed text. Further studies on a larger dataset are needed to develop and to test new sets of product-oriented idiolectal markers, which will contribute to our understanding of an idiolect and enhance development of new methods for idiolect identification.

Keywords: Authorship Attribution, Corpus Linguistics, Linguistic Resource, Idiolect, Typing Behavior, Multivariate Data Analysis.

1 Introduction

Authorship attribution (AA), i.e. the problem of attributing a text to its author, is a task of a considerable practical importance. In recent years, this problem has been tackled mostly as one of the text classification problems and one of the subtasks of user identification problem. Different types of linguistics markers have been used as features (word and part-of-speech statistics, text length, etc.) [7]. Since the main ob-

ject of modern AA is a typed text, another line of research is a study of the speed and rhythm of typing patterns (keystroke dynamics, KD) [9]. Comparative studies of stylometric and typing features for AA report significantly higher accuracies of KD-based models despite being two orders of magnitude smaller [15] (see also [17] for similar conclusions on superiority of KD feature over stylometric markers in AA task). The disadvantage of KD-based features is a lack on interpretability since no linguistic information is typically used in such analysis. Moreover, as a rule, studies on KD are based on texts produced in the same communication situation (single-domain user identification) without considering the stability of idiolectal markers in cross-domain scenario, with rare exceptions showing sensitivity of KD markers to small differences in writing tasks [4]. The same is also true for stylometric features: most of them were shown to be domain dependent [10], which explains the extreme difficulty of cross-domain AA.

Since despite decades of research AA remains challenging, especially in a typical forensic scenario characterized by a small number of samples (texts), heterogeneity of training and testing documents in terms of genre, topic, mode, etc. [14] and demand for interpretability of the results, new types of idiolectal markers are urgently needed which would combine both typing data and linguistic information and as well as research on their stability under the effect of different factors of intraidiolectal variation (psychological state of the author, genre of the text, mode, etc.). This paper aims to make the first steps in this direction. The contribution of the paper is as follows.

1. The first resource to study process-oriented idiolectal features in Russian typed texts is introduced.
2. State-of-the-art methods from bioinformatics designed for multivariate data analysis involving a small sample size and a large number of highly correlated predictors have been applied for the first time to AA task and related text type effect elimination.
3. For the first time discriminative ability of process-oriented idiolectal features has been researched for the task of AA of highly heterogeneous texts.

The rest of the paper is organized as follows. In Section 2, related work on process-oriented idiolectal features and their usefulness in idiolectal research is briefly outlined. Section 3 describes the methodology of the study. First, the research corpus is introduced, the process of its compilation and annotation is described. Second, two types of process-oriented idiolectal features are presented. Third, the methods of multivariate data analysis are described. In Section 4, the results of the pilot study on AA of highly heterogeneous texts based on process-oriented idiolectal features are presented. In Section 5 the conclusions are drawn, and directions for future work are outlined.

2 Related Work

Although there are a lot of works dealing with AA using stylometry and keystroke information separately, there are only few ones which use process-oriented idiolectal features.

Hybrid, process-oriented features for AA was introduced for the first time in [14]. They focused on features based on bursts - cognitive units of texts production between two pauses of fixed length. Their best performing burst feature subset had 72 features comprising of word creation, lexical complexity, revising style, keyboard proficiency, and pre- and post-pause features. No further linguistic analysis of the features is provided, though.

Later the same team applied "language production" features to classify authors by group (authorship profiling) [2]. This time feature set included part-of-speech (POS) pauses (mean length of a pause before and after each word as represented by its part of speech), the pause time before and after punctuation marks, misspelling pauses, revision features and typing burst features. One of interesting implications of the study is that some differences between female and male texts in language production and keystroke dynamics were revealed, but not in traditional stylometric indicators.

Overall, the use of process-oriented features in AA, despite having been shown to be prospective, is heavily understudied, particularly in a cross-domain scenario. This is partially explained by the difficulties in collecting appropriate datasets. To the best of our knowledge, to date there is only one publicly available dataset with annotation for writing events (namely revisions) [5].

3 Methodology

3.1 Dataset

To study the process-oriented idiolectal markers and their usefulness for AA, the dataset "Multifactor" created in Corpus Idiolectology Lab was used. This dataset contains highly heterogeneous texts produced by the same authors. Texts differ in modes (oral and written), types (dialogues and monologues), genres, topics. Along with texts, dataset contains sound files, transcriptions of oral texts manually annotated for elementary discourse units (EDUs) as well as video files containing screen recording of written text production. The dataset consists of 242 texts (2 oral dialogues, 3 written dialogues, 3 oral monologues, 3 written monologues) by 22 authors (11 females) of age group 18–24. The dataset allows us to conduct research on the stability of idiolectal markers in different domains.

Currently, the corpus is available by request, but there is an ongoing work on preparing it for public use.

For this particular study, three authors were selected randomly from Multifactor corpus (ID 1, female, ID 2, female, ID 3, male). Each author contributed 3 dialogues (with different interlocutors; topic – "Guess the movie on its description") and 3 monologue texts (letter of complaint, essay, short movie retelling). Manual annotation

of the timing of text production has been performed by two linguists with the use of ELAN software [6]. All the pauses (time with no action longer than 200 ms) were identified on time scale, as well as all actions (writing events between pauses) (see Fig. 1). After ELAN timing annotation for each event, manual annotation was done as follows. The pauses were classified depending on their duration, s; PAUSE 1 (< 0.49); PAUSE 2 ([0.5; 0.99]), PAUSE 3 ([1; 1.99]), PAUSE 4 ([2; 5]), PAUSE 5 (> 5). These thresholds were selected based on writing research literature (see [12] and references therein). "Full" words (as well as words with one or two misspellings allowing to unambiguously detect POS) were supplied with their POS tags, punctuation marks were assigned their labels (PERIOD, COMMA, DASH, QM (for question mark), EM (for exclamation mark)), deleted words were replaced with DEL tag, parts of words were replaced with PART tags, corrections were replaced with CORR tags. For dialogues additional tags were used: BREAK – start of a new line in one turn, TURN – the end of a turn.

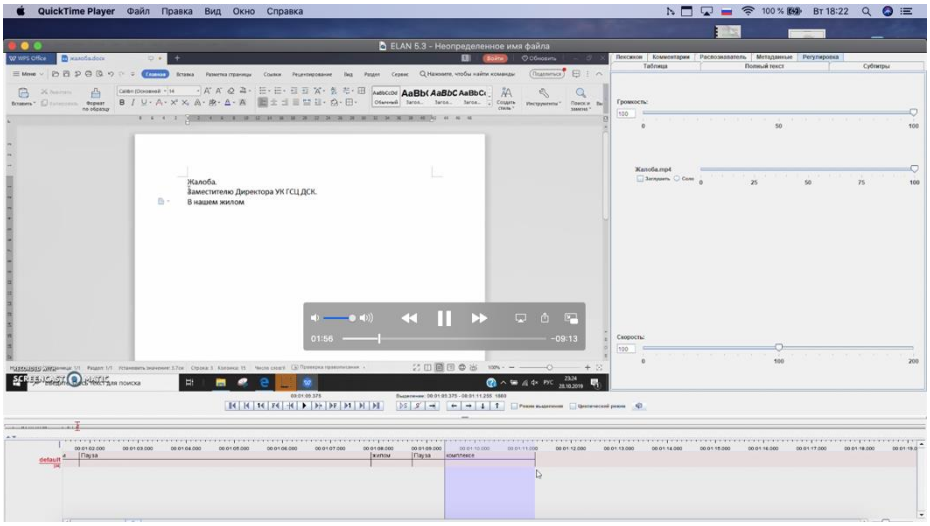

Fig. 1. ELAN annotation window

An example of annotation of the text "Привет/ ды ниче / валяюсь / а ты / ?" ("Hi / I'm ok / chilling / what about you / ?") where slash means the end of the line is shown below (Table 1).

Table 1. An example of text annotation

Event ID	Beginning time	Ending time	Duration	ELAN	Annotation
1	00:00:00.000	00:00:04.440	00:00:04.440	Pause	PAUSE4
2	00:00:04.440	00:00:06.110	00:00:01.670	Привет	INT
3	00:00:06.110	00:00:13.260	00:00:07.150	Pause	TURN
4	00:00:13.260	00:00:13.630	00:00:00.370	ды	PARTICLE
5	00:00:13.630	00:00:13.970	00:00:00.340	Pause	PAUSE1
6	00:00:13.970	00:00:14.380	00:00:00.410	ни	PART
7	00:00:14.380	00:00:15.430	00:00:01.050	Pause	PAUSE2
8	00:00:15.430	00:00:15.635	00:00:00.205	ч	PART
9	00:00:15.635	00:00:16.845	00:00:01.210	Pause	PAUSE3
10	00:00:16.845	00:00:17.000	00:00:00.155	е	PART
11	00:00:17.000	00:00:19.300	00:00:02.300	Pause	BREAK
12	00:00:19.300	00:00:20.795	00:00:01.495	валяюм	VERB
13	00:00:20.795	00:00:21.645	00:00:00.850	Pause	PAUSE2
14	00:00:21.645	00:00:23.195	00:00:01.550	~~валяюм~~	DEL
15	00:00:23.195	00:00:25.370	00:00:02.175	Pause	BREAK
16	00:00:25.370	00:00:26.880	00:00:01.510	валяюсь	VERB
17	00:00:26.880	00:00:28.195	00:00:01.315	Pause	BREAK
18	00:00:28.195	00:00:28.355	00:00:00.160	а	PARTICLE
19	00:00:28.355	00:00:28.730	00:00:00.375	Pause	PAUSE1
20	00:00:28.730	00:00:29.170	00:00:00.440	ты	PRON

To avoid the effect of text length and total timing differences between authors which could spur our results, we restricted ourselves with the first 400 events for each text. The mean text length of final texts produced during these 400 events was 147 words (SD = 22 words) for dialogues and 123 words (SD = 30) for monologues, the latter being shorter than the former (paired t-test $p = 0.0008$). We calculated the mean time spent in each POS tags, DEL and PART by summing durations of these states in text and dividing it by the number of states in each text. The normality of the data distribution was tested using the Shapiro–Wilk test but not confirmed. The results of non-parametric Wilcoxon signed-rank test with FDR correction showed no differences between the mean time spent in each state except for PART (p.adjust = 0.03906): in dialogues mean time spent in PART is higher.

3.2 Feature description

Two types of process-oriented idiolectal markers were used in this study. The first group of markers characterize non-sequential characteristics of text production process. We name them general production features (GPF).

1. General production features:

- **Pure_Words** = total words / (total time – pause time – part_del_corr). This is a measure for author productivity expressing the ratio of words to time spent in productive writing.
- **Corr_words** = part_del_corr / total words. The ratio of revision events to total words.
- **WL (word length)**: total words/total char. The mean word length in characters.
- **PM_PAUSE:** the ratio of time spent in punctuation marks to time spent in pauses.
- **WL_Time_thinking:** the ratio of the mean word length to time spent in editing and pausing.
- **VERB_DEL –** the ratio of number of words to the number of deletions.
- **PM_DEL –** the ratio of number of punctuation marks to the number of deletions.
- **NOUN_PART, VERB_PART, ADV_PART, PREP_PART, PM_PART** – the ratio of number of nouns, verbs, adverbs, prepositions, punctuation marks to the number of PART events.
- **PREP_CONJ** – ratio of total time spent in prepositions to the total time spent in conjunctions.
- **PART_time, PREP_CONJ, CONJ_time, VERB_time, ADV_time, COMMA_time, PREP_time, DEL_time** – total time spent in prepositions, conjunctions, verbs, adverbs, comma, DEL, correspondingly.

2. Sequential features (BIGRAM*)*

Sequential features – the frequencies of all pairs of adjacent events (i.e. bigrams) – were calculated providing one of them is PAUSE event, which resulted in 164 features. A pre-filtering step to remove bigrams for which the sum of counts is below a certain threshold (we chose 1 % cut-off) compared to the total sum of all counts was made1. Thus, the original features set was reduced to 64.

Prior to the main analysis of each feature set, we applied Hellinger standardization where each element is divided by its row sum, after the square root of each element is calculated. This type of standardization was selected since it yields low weights to variables with low counts and a lot of zeros [3].

3.3 Methods

For this pilot study different methods for identifying the structure in multivariate data were used as implemented in R package mixOmics (for the sake of brevity we do not

[1] The function for prefiltering was borrowed from http://mixomics.org/mixmc/mixmc-preprocessing/ (assessed 25/10/2020)

provide description of the methods here and refer reader to [16]). Namely, Principal component analysis (PCA) and multilevel PCA (PCA with extraction of the within variation matrix) were used to visualize the idiolectal markers variation according to the text type (monologue and dialogue) and author. The multilevel function first decomposes the within (text type) from the between (author) variance in the data sets. This is crucial for our task, since we hypothesize a strong text-type effect in our data which could complicate AA. To complement PCA results, we also applied variation partitioning analysis using R package vegan (see [18] for more details).

Partial Least Squares Discriminant Analysis (PLS-DA) was used for text class (author ID) prediction. Stratified 2-fold cross-validation as implemented in mixOmics (since mixOmics package only allows 'n of classes – 1' number of folds) and 6-fold (default value) cross-validation as implemented in R package **RVAideMemoire** [8] was applied to assess the performance of PLS-DA models as well as leave-one-out cross-validation, but the results were similar and we report only the results of 6-fold cross-validation. We also performed permutation tests to assess the significance of PLS-DA models using R package RVAideMemoire. Sample classifiers were scrambled 999 times and models re-calculated to establish the likelihood of achieving the same result by chance. Then, the dataset integration method DIABLO was used to assess the predictive ability of combination of two feature sets with respect to the outcome (**Author**). This type of N-integration analysis (i.e. suitable for research where several features set measured on the same individuals – the same N – are available, which is the case for most idiolectal studies) developed by the mixOmics team is aimed to identify a highly correlated multi-dataset signature discriminating the samples (**texts**) in accordance with their group (**author**).

MixOmics is well-documented R package which allows one to perform different types of state-of-the-art analysis on multivariate data with a special focus on variable selection and data visualization. Originally designed for "omics" (large-scale biological datasets) data, these methods are suitable for idiolect studies since idiolectal data and idiolect identification tasks have a lot in common with omics data and related problems: multicollinearity and a large number of predictors ($p > N$ problem), a small number of samples, strong connections between different datasets (linguistic levels of idiolect) describing the same individuals. Moreover, using the above mentioned methods with a strong focus on variable importance, identification of the sources of variability and data visualization could shift the paradigm in AA studies most of which are currently using a purely engineering approach with focus on prediction rather than interpretability.

4 Results

First, we performed an experiment with general production features (GPF). Broken-stick model [18] has shown that the first two components are worth being interpreted. They account for 65 % of variation. As Fig. 2 shows, there is a clear tendency of text type-based (dialogue (denoted as D_) versus monologue (M_)) clustering over the comp 1 which explains the largest part of variation (49 %). Variation partitioning

analysis has proved PCA findings about a text type as the main source of variation in our data. Adj.R.squared for the factor “Text type” is 0.24 (i.e. type explains 24 % of variance in our data from 65 % that could be possible explained by the first 2 components), while “Author” only 0.06 (i.e. author explains only 6 % of variance in the data). A permutation test has shown that the results of the analysis is significant both for the model with two factors (p=0.002) and for the model with a text type as the main factor (p= 0.001) and author as covariate and for author as main factor and text type as covariate (p = 0.027).

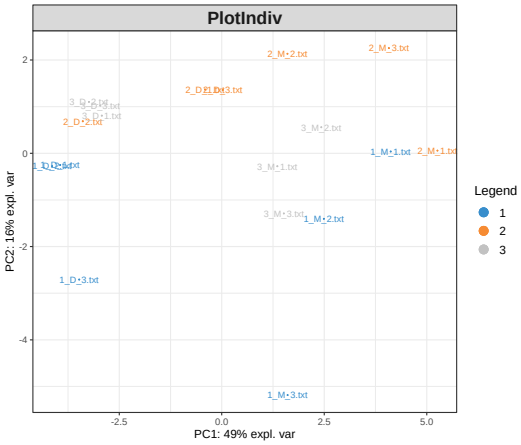

Fig. 2. PCA on GPF feature set

However, when we applied multilevel PCA for text-type effect elimination, the tendency to author-based clustering has been revealed (Fig. 3).

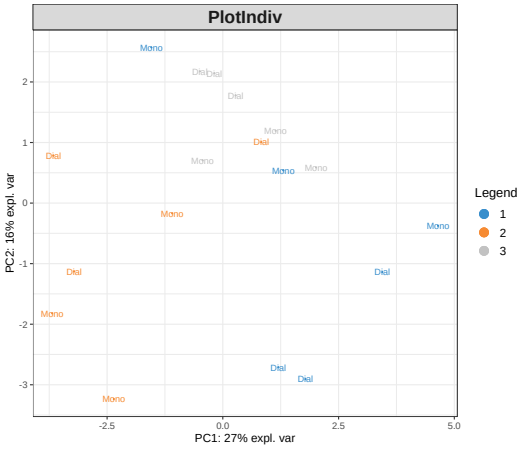

Fig. 3. Multilevel PCA with genre effect elimination for the GPF feature set

Having revealed the main sources of variation in our data using an unsupervised approach, we next move on to a classification experiment to assess the predictive ability of variables. As expected, PLS-DA error rate on data without variance decomposition is high (0.42778), although the model is still significant (permutation test p-value = 0.034). Pairwise permutation tests with FDR correction revealed that all authors differ from each one (all $p < 0.05$). However, when we constructed PLS-DA on data with the eliminated text-type effect, error rate decreased to 0.175. The permutation test has shown that results are statistically significant ($p = 0.001$; all pairwise permutation tests $p < 0.05$).

For visualization of the results of PLS-DA model we used color-coded Clustered Image Maps (CIMs) ("heat maps"). CIM is a 2-dimensional visualization of a real-valued matrix with rows and columns reordered according to some hierarchical clustering method (we used Ward method and Euclidean distance) to identify some interesting patterns in data (i.e. simultaneous clustering of samples and variables). CIM (Fig. 4) shows 3 clear authorial groups (texts are rows) as well as 2 large clusters of authorial markers (variables are columns): one cluster contains variables expressing general productivity; second cluster consists of variables reflecting different particular characteristics of writing processes which are further divided into two smaller clusters.

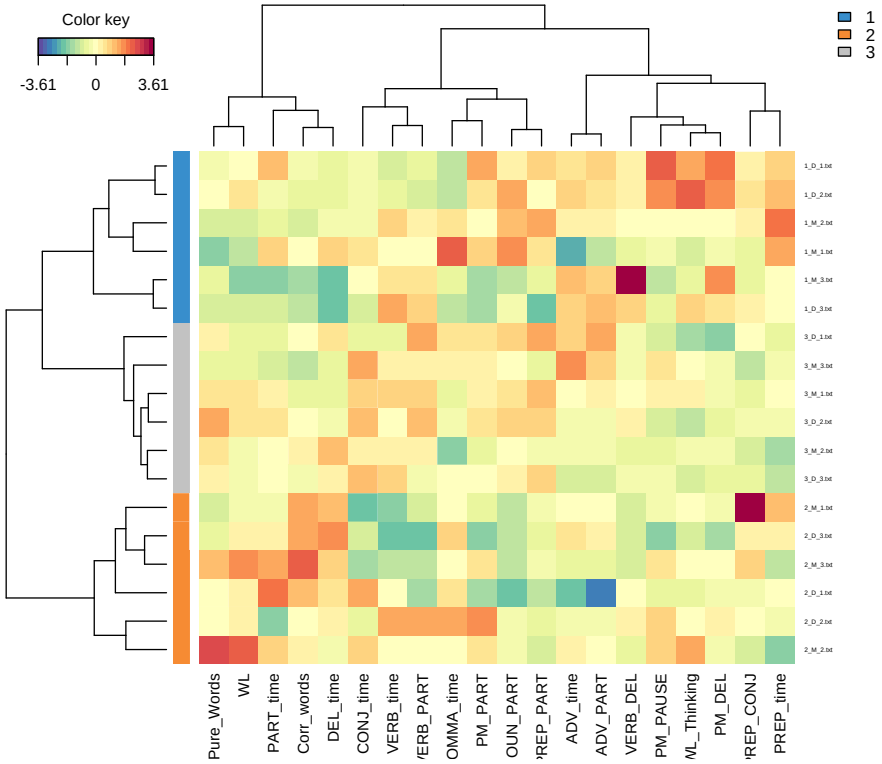

Fig. 4. Clustered Image Map on multilevel PLS-DA, GPF set

Broken-stick criterion for PCA with BIGRAM features shows that the first 2 components are meaningful, which accounts for 45 % of variation. Fig. 5 shows that a text type is the main source of variation. Multilevel PCA eliminates a text-type effect: clear authorial clusters are revealed (Fig. 6).

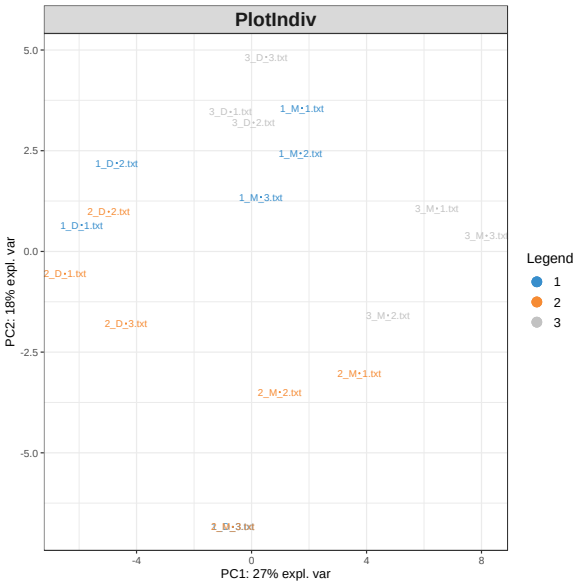

Fig. 5. PCA on the BIGRAM feature set

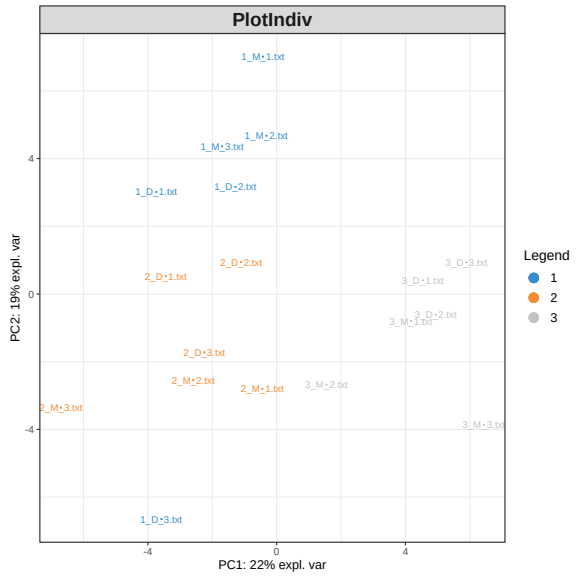

Fig. 6. Multilevel PCA on the BIGRAM feature set

Variation partition analysis shows that both factors contribute to variation: Adj.R.squared for the factor "Text type" is 0.14 (i.e. it explains 14 % of variance in our data from 45 % that could be possible explained by the first 2 components), while Adj.R.squared for "Author" is 0.12 (i.e. it accounts for 12 % of variance in the data). The permutation test has shown that the results of the analysis are significant both for the model with two factors (p=0.001) and for that with one factor as the main and another as the covariate one (p= 0.001 for both models).

PLS-DA performed without variance decomposition shows ER = 0.15 (p-value = 0.001), ER of the model on data with variance decomposition is 0.086111 (p-value = 0.001). For both models, $p < 0.05$ in pairwise comparisons, except for the model without variance decomposition ($p = 0.056$ for pair "Author1 – Author 2").

CIM revealed 3 author-based clusters with 1 incorrectly classified text by Author 1 (Fig. 7) as well as 2 large clusters of features. One cluster consists of the features reflecting pause behavior in revisions, the other one contains features reflecting particular characteristics of pausing behavior.

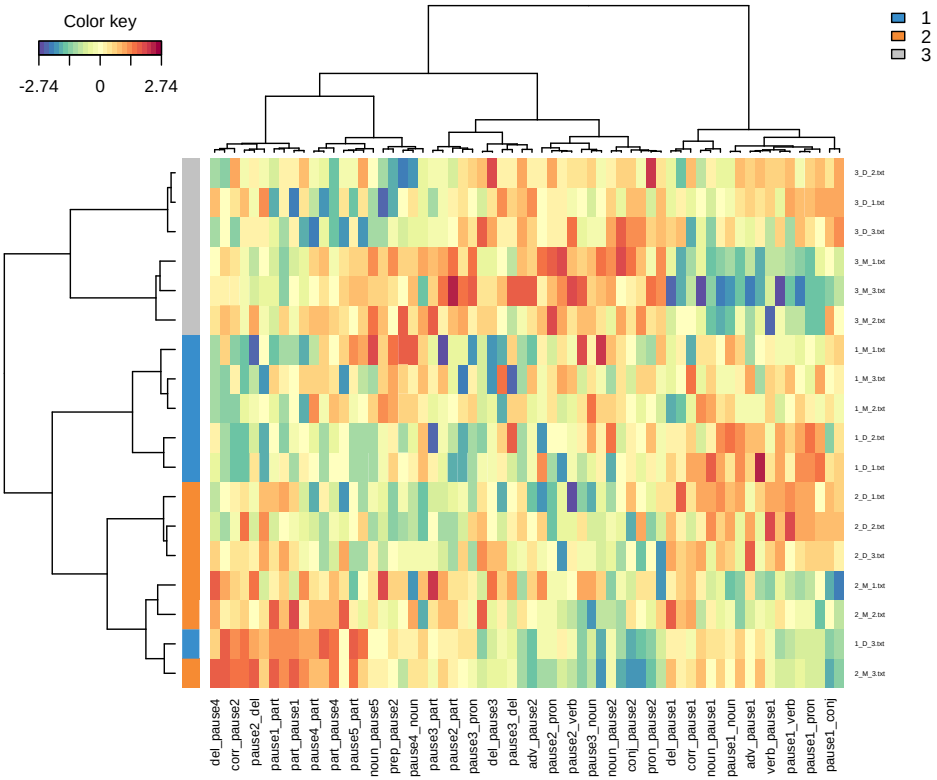

Fig. 7. CIM on multilevel PLS-DA performed on BIGRAM feature set

Next we move on to the feature set integration (two-block DIABLO). The permutation test based on cross-validation confirmed the significance of the DIABLO model (ER = 0.14167, p-value = 0.001). The first components from each data set are highly correlated (0.81) to each other; the same is true for the second components (0.82).

Fig. 8 displays the variables from two blocks selected on component 1 and 2 (cut-off = 0.5). The clusters of points indicate a strong correlation between the variables.

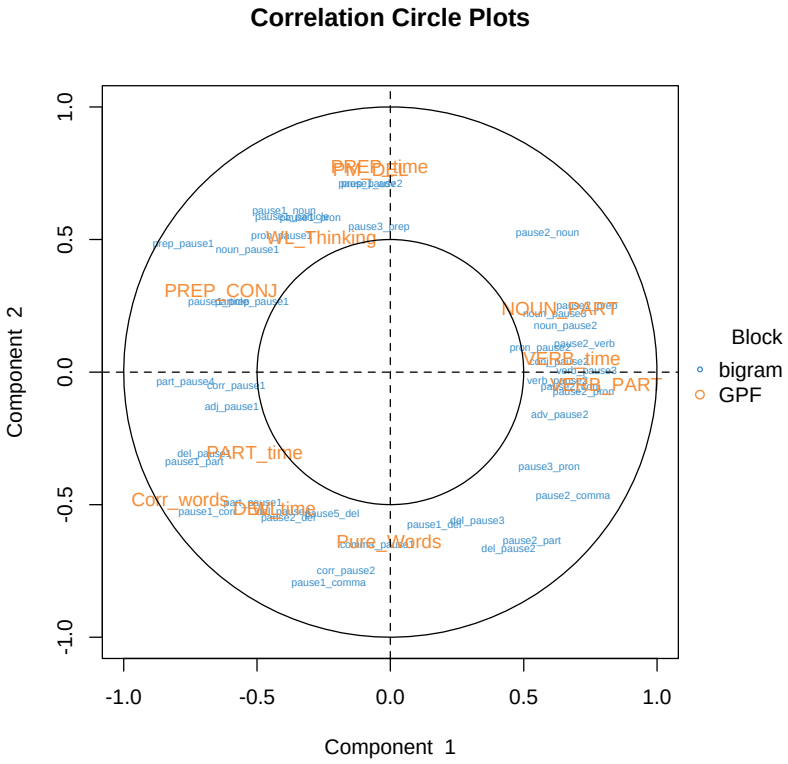

Fig. 8. Correlation circle for two feature sets

CIM on DIABLO model (Fig. 9) allows us to see the signature of each author over the two feature sets. Irrespective of a text type, Author 1 is characterized by the higher values of NOUN_PART, Prep_CONJ, WL_Thinking, PM__DEL, prep_pause1, prep_pause2, prep_time, PM_DEL, VERB_DEL, pause3_prep, pause1_adv, adv_time, adv_part, part_pause2, verb_pause1, pause2_corr. Overall, author 1 is typical spent typing time in prepositions and adverbs, as well as verbs and less time spent in revisions.

Author 2 is characterized by higher values of corr_words, higher WL, PART_time, DEL_time, del_pause1, pause1_part, adj_pause1, PM_pause, comma_pause1, pause1_comma, pause2_del, del_pause4. Most of the features which characterize this author are related to the general productivity (she spends more time in revisions, although has higher word complexity as assessed by word length) and punctuation behavior. The only POS feature is short pause + adjective bigram.

Author 3 is characterized by a higher mean time spent in CONJ, higher values of PREP_PART, VERB_part, VERB_time, PM_PART, verb_pause3, pause2_prep, pause2_verb, pause3_pron, PURE_WORDS, pause1_del, corr_pause2, del_pause2, del_pause3, pause3_noun, pause2_comma. Author 3 is characterized by a larger amount of time spent in conjunctions, verbs, less time spent in PART, i.e. higher speed of word retrieval.

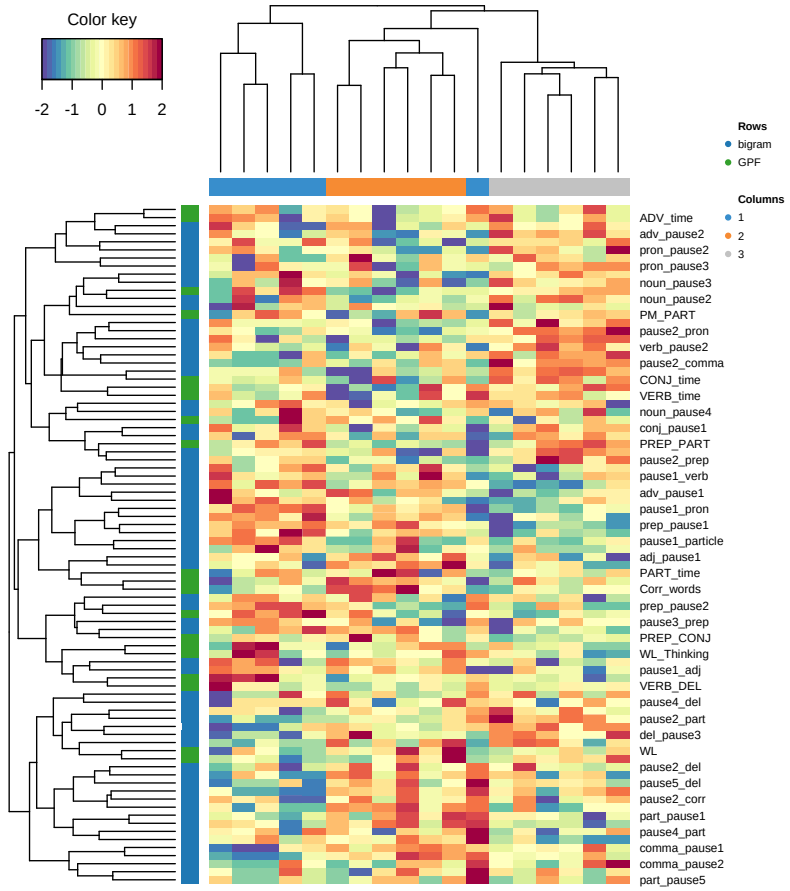

Fig. 9. CIM on DIABLO model

Here again we see two large clusters of variables, one of them expressing general characteristics of productivity (at the bottom of Fig. 9), while the second one expresses particular characteristics of the writing process.

5 Conclusions and Future Work

We have performed a pilot study into the stability and discriminative ability of a process-oriented set of idiolectal markers which combine information from stylometry and keystroke dynamics in a very complicated heterogeneous (dialogue and monologue texts in dataset) authorship attribution scenario.

Based on the performed experiments, we have arrived at the following conclusions:

1. Process-oriented idiolectal markers could be used to detect authors in a highly heterogeneous dataset.
2. Process-oriented idiolectal markers reflecting information on the duration on pauses before and after words with a known morphological class (POS) as well as a writing process event related to revision are better predictors of author than process-oriented idiolectal markers expressing non-sequential information.
3. State-of-the-art methods of analysis and visualization of multivariate data developed originally for large biological datasets are suitable for analyzing idiolects since biological and linguistic data have a lot in common.
4. The methods for variance decomposition allow one to eliminate a strong text-type effect and could be used in cross-domain authorship attribution, which is the most complicated type of AA tasks.

As with every pilot study, the presented results could not be generalized, however, they definitely point out the ways of future work.

1. It is necessary to expand the corpus of texts annotated for a process-oriented set of idiolectal markers. In order to do it in a more efficient and a less-time-consuming manner, it is necessary to develop methods which could extract these features automatically.
2. It is urgent to broaden the set of a process-oriented idiolectal markers and a range of communication tasks authors are involved in.
3. One of the prospective ways of future research is to assess the effect of the medium of text product (pen – physical keyboard – touch screen keyboard) on stability and discriminative ability of process-oriented idiolectal markers.

Acknowledgement

This work is supported by the grant No. 18-78-10081 from Russian Science Foundation, which is gratefully acknowledged.

References

1. Belman, A. K. et al: Insights from BB-MAS – A Large Dataset for Typing, Gait and Swipes of the Same Person on Desktop, Tablet and Phone (2019). arXiv:abs/1912.02736 (2019).
2. Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., Rosenberg, A.: Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. In: International Journal of Human-Computer Studies, 82, 57–68 (2015).
3. Buttigieg, P.L., Ramette, A.: A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. In: FEMS Microbiol Ecol., 90, 543–550 (2014).
4. Conijn, R., Jens, R., van Zaanen, M.: Understanding the keystroke log: the effect of writing task on keystroke features. In: Reading and Writing, 32(9), 2353–2374 (2019).
5. Conijn, R., Speltz, E. D., van Zaanen, M., van Waes, L., Chukharev-Hudilainen, E.: A product and process oriented tagset for revisions in writing. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 363–368 (2020).
6. ELAN (Version 5.9) The Language Archive, https://archive.mpi.nl/tla/elan, last accessed 2020/03/01.
7. Grieve, J.: Quantitative authorship attribution: an evaluation of techniques. Literary and Linguistic Computing, 22 (3), 251–270 (2007).
8. RVAideMemoire: testing and plotting procedures for biostatistics. R Package Version 0.9-69, https://CRAN.R-project.org/package=RVAideMemoire, last accessed 2020/02/17.
9. Kostyuchenko, E., Kadyrov, R.: Influence of the System Parameters on the Final Accuracy for the User Identification by Free-Text Keystroke Dynamics. In: Bhatia, S., Tiwari, S., Mishra, K., Trivedi, M. (eds.) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, 759. Springer, Singapore (2019).
10. Litvinova, T.: Stylometrics Features Under Domain Shift: Do They Really "Context-Independent"? In: Karpov, A., Potapova, R. (eds.) Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science, 12335. Springer, Cham (2020).
11. Locklear, H., et al: Continuous authentication with cognition-centric text production and revision features. In: IEEE International Joint Conference on Biometrics, 1–8. Clearwater, FL (2014).
12. Medimorec, S., Risko, E.F.: Pauses in written composition: on the importance of where writers pause. In: Reading and Writing, 30, 1267–1285 (2017).
13. Monaco, J., Stewart, J., Cha, S.-H, Tappert, C.: Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In: 2013 IEEE Sixth Intl. Conf. on Biometrics: Theory, Applications and Systems, 1–8 (2013).
14. Panicheva, P., Litvinova, T.: Authorship Attribution in Russian in Real-World Forensics Scenario. In: Martín-Vide, C., Purver, M., Pollak, S. (eds.). Statistical Language and Speech Processing. SLSP 2019, Lecture Notes in Computer Science, 11816. Springer, Cham (2019).
15. Plank, B.: Predicting authorship and author traits from keystroke dynamics. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, 98–104 (2018).
16. Rohart, F., Gautier, B., Singh, A., Lê Cao, K-A.: mixOmics: an R package for 'omics' feature selection and multiple data integration. PLoS Comput Biol, 13(11) (2017).

17. Stewart, J. C., Monaco, J. V., Cha, S. H., Tappert, C. C.: An investigation of keystroke and stylometry traits for authenticating online test takers. In: 2011 International Joint Conference on Biometrics (IJCB), 1–7 (2011).
18. Analysis of community ecology data in R, https://www.davidzeleny.net/anadat-r/doku.php/en:start, last accessed 2020/01/20.