Authorship Attribution and Author Profiling of Lithuanian Literary Texts

Jurgita Kapočiūtė-Dzikienė
Vytautas Magnus University
K. Donelaičio 58, LT-44248,
Kaunas, Lithuania
jurgita.k.dz@gmail.com

Andrius Utka
Vytautas Magnus University
K. Donelaičio 58, LT-44248,
Kaunas, Lithuania
a.utka@hmf.vdu.lt

Ligita Šarkutė
Kaunas Univ. of Technology
K. Donelaičio 73, LT-44029,
Kaunas, Lithuania
ligita.sarkute@ktu.lt

Abstract

In this work we are solving authorship attribution and author profiling tasks (by focusing on the age and gender dimensions) for the Lithuanian language. This paper reports the first results on literary texts, which we compared to the results, previously obtained with different functional styles and language types (i.e., parliamentary transcripts and forum posts).

Using the Naïve Bayes Multinomial and Support Vector Machine methods we investigated an impact of various stylistic, character, lexical, morpho-syntactic features, and their combinations; the different author set sizes of 3, 5, 10, 20, 50, and 100 candidate authors; and the dataset sizes of 100, 300, 500, 1,000, 2,000, and 5,000 instances in each class. The highest 89.2% accuracy in the authorship attribution task using a maximum number of candidate authors was achieved with the Naïve Bayes Multinomial method and document-level character tri-grams. The highest 78.3% accuracy in the author profiling task focusing on the age dimension was achieved with the Support Vector Machine method and token lemmas. An accuracy reached 100% in the author profiling task focusing on the gender dimension with the Naïve Bayes Multinomial method and rather small datasets, where various lexical, morpho-syntactic, and character feature types demonstrated a very similar performance.

1 Introduction

With the constant influx of anonymous or pseudonymous electronic text documents (forum posts, Internet comments, tweets, etc.) the authorship analysis is becoming more and more topical. In this respect it is important to consider the anonymity factor, as it allows everyone to express their opinions freely, but on the other hand, opens a gate for different cyber-crimes. Therefore the authorship research –which for a long time in the past was mainly focused on literary questions of unknown or disputed authorship– drifts towards more practical applications in such domains as forensics, security, user targeted services, etc. Available text corpora, linguistic tools, and sophisticated methods even more accelerate the development of the authorship research field, which is no longer limited to authorship attribution (when identifying who, from a closed-set of candidate authors, is the actual author of a given anonymous text document) only. Other research directions involve the author verification task (when deciding if a given text is written by a certain author or not); the plagiarism detection task (when searching for similarities between two different texts or parts within a single text); the author profiling task (when extracting information about author's characteristics, typically covering the basic demographic dimensions as the age, gender, native language or psychometric traits); etc. In this paper we focus on authorship attribution (AA) and author profiling (AP) problems covering the age and gender dimensions.

Some researchers claim that in the scenario when an author makes no efforts to modify his/her writing style, authorship identification problems can be tackled due to an existing "stylometric fingerprint" notion –an individual and uncontrolled habit to express thoughts in certain unique ways, which is kept constant in all writings by the same author. Van Halteren (2005) even named this phenomenon a "human stylome" in analogy to a DNA "genome". However, Juola (2007) argues that strict implications are not absolutely correct, be-

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 96–105,
Hissar, Bulgaria, 10–11 September 2015.

cause the "genome" is stable, but the writing style tends to evolve over time. More stable (e.g., gender) and changing (e.g., age, social status, education, etc.) demographic characteristics affect the writing style, thus making AP a solvable task for these dimensions.

With the breakthrough of the Internet era literary texts in the authorship research were gradually replaced with e-mails (Abbasi and Chen, 2008), (de Vel et al., 2001), web forum messages (Solorio et al., 2011), online chats (Cristani et al., 2012), (Inches et al., 2013), Internet blogs (Koppel et al., 2011) or tweets (Sousa-Silva et al., 2011; Schwartz et al., 2013), which, in turn, contributed to a development of Computational Linguistic methods able to cope effectively with the following problems: short texts, non-normative language texts, many candidate authors, etc. The discovered advanced techniques helped to achieve even higher accuracy in AA and AP tasks on literary texts (under so-called "ideal conditions"). Although the research on literary texts have lost popularity due to the decrease in demand of their practical applications, the results obtained on literary texts can still be interesting from the scientific point of view, as it may perform some kind of a baseline function in the comparative research. In this respect we believe that our present paper, which focuses on literary texts, will deliver valuable results. Besides, the obtained AA and AP results will be compared with the results previously reported on the Lithuanian language covering other functional styles and language types.

2 Related Works

Despite archaic rule-based approaches (attributing texts to authors/characteristics depending on a set of manually constructed rules) and some rare attempts to deal with unlabeled data (Nasir et al., 2014; Qian et al., 2014) automatic AA and AP tasks are tackled with Supervised Machine Learning (SML) or Similarity-Based (SB) techniques (for review see (Stamatatos, 2009)). In the SML paradigm, texts of known authorship/characteristic (training data) are used to construct a classifier which afterwards attributes anonymous documents. In the SB paradigm, an anonymous text is attributed to the particular author/characteristic whose text is the most similar according to some calculated similarity measure. The comparative experiments prove superiority of the SB methods over the SML techniques, e.g., Memory-Based Learning produced better results compared to Naïve Bayes and Decision Trees (Zhao and Zobel, 2005); the Delta method surpassed performance levels achieved by the popular Support Vector Machine method (Jockers and Witten, 2010). However, the SB approaches are considered to be more suitable for the problems with a big number of classes and limited training data, e.g., the Memory-Based Learning method applied on 145 authors outperformed Support Vector Machines (Luyckx and Daelemans, 2008), applied on 100,000 candidate authors outperformed Naïve Bayes, Support Vector Machine and Regularized Least Squares Classification (Narayanan et al., 2012). In our research we have at most 100 candidate authors, 6 age and 2 gender groups, therefore the SML approaches seem to be the most suitable choice. Besides, many AA and AP tasks are solved using the popular Support Vector Machine method, which in the contemporary computational research is considered as the most accurate, thus the most suitable technique for different text classification problems (e.g., superiority of Support Vector Machine is proved in (Zheng et al., 2006)). However, a selection of classification method itself is not as important as a proper selection of an appropriate feature type.

Starting from Mendenhall (1887) the first stylometric techniques were based on the quantitative features (so-called "style markers") such as a sentence or word length, number of syllables per word, type-token ratio, vocabulary richness function, lexical repetition, etc. (for review see (Holmes, 1998)). However, these feature types are considered to be suitable only for homogeneous and long texts (e.g., entire books) and for the datasets having only a few candidate authors. The first modern pioneering work of Mosteller and Wallace (1963) –who obtained promising AA results on The Federalist papers with the Bayesian method applied on frequencies of a few dozens function words– triggered many posterior experiments with various feature types. In the contemporary research the most widespread approach is to represent text documents as vectors of frequencies, which elements cover specific layers of linguistic information (lexical, morpho-syntactic, semantic, character, etc.). The best feature types are determined only after an experimental investigation.

Since most AA and AP research works deal with Germanic languages, providing no recommendations that could work with the morphologically rich, highly inflective, derivationally complex languages (such as Lithuanian), having relatively free word order in sentences, our focus is on the research done for the Baltic and Slavic languages, which by their nature and characteristics are the most similar to Lithuanian.

The AA experiments for the Polish language were performed with the literary texts of 2 authors using the feed-forward multilayer Perceptron method (with one or two hidden layers) and the sigmoid activation function trained by the back-propagation algorithm (Stańczyk and Cyran, 2007). The experiments with lexical (function words), syntactic (punctuation marks), and combination of both feature types revealed superiority of syntactic features. Eder (2011) applied the Delta method on the Polish, English, Latin, German datasets each containing 20 prose writers and then compared obtained AA results. A bootstrap-like procedure –testing a large number of randomly chosen permutations of original data with the k-Nearest Neighbor method in each trial and calculating an average accuracy score– helped to avoid fuzziness with unconvincing results. The best results for the Polish language texts were achieved with a mix of word unigrams and bigrams, with word unigrams for English, with a combination of words and character penta-grams for Latin, with character tri-grams for German.

Kukushkina et al. (2001) applied first-order Markov chains on the Russian literary texts written by 82 authors. All matrices –containing transition frequency pairs of text elements– composed during the training process for each candidate author were later used to compute probabilities of anonymous texts. The researchers investigated word-level (an original word form or it's lemma) character bigrams, pairs of coarse-grained or fine-grained part-of-speech tags and obtained the best results with word-level (in the original form) character bigrams. Kanishcheva (2014) presented the implemented software able to solve AA tasks for the Russian language. The offered linguistic model is based on statistical characteristics and can fill the lexical database of the author's vocabulary. Any attribution decision is taken after calculations of a proximity value between texts.

For the Croatian language the AA task was solved using the Support Vector Machine method with the radial basis (Reicher et al., 2010). The researchers used 4 datasets (newspaper texts of 25 authors, on-line blogs of 22 authors, Croatian literature classics of 20 authors, Internet forum posts of 19 authors). They tested a big variety of features and their combinations: function words, idf weighted function words, frequencies of coarse-grained part-of-speech tags, fine-grained part-of-speech tags with normalized frequencies, part-of-speech tri-grams, part-of-speech tri-grams with function words, other features (including punctuation, frequencies of word lengths, sentence-length frequency values, etc.). The best results were achieved with a combination of function words, punctuation marks, word and sentence length frequency values.

Zečević (2011) investigated byte-level character n-grams on the Serbian newspaper dataset of 3 authors. The researcher explored an influence of the author profile size (varying from 20 up to 5,000 most frequent n-grams) and the n-gram length (up to 7). All n-grams were stored in a structure called a prefix tree; an author attribution decision was taken by the 1-Nearest Neighbor algorithm based on the distance metric combining the dissimilarity measure and the simplified profile intersection. The best results were achieved with the n-grams of $n > 2$ and the profile size larger than 500. In the posterior work (Zečević and Utvić, 2012) researchers added 3 more candidate authors to the dataset and investigated an impact of syllables using the simplified profile intersection similarity measure. However, syllables were not robust enough to outperform byte-level character n-grams.

Other research works (as for the Slovene language in (Zwitter Vitez, 2012)) demonstrate potentials to solve AA or AP tasks. They represent available text corpora, linguistic tools and discuss possible methods, feature types, an importance of AA and AP tasks, etc.

For the Lithuanian language the AA research was done with 100 candidate authors and two datasets of parliamentary transcripts and forum posts (Kapočiūtė-Dzikienė et al., 2015). The researchers explored the Naïve Bayes Multinomial and Support Vector Machine methods with a big variety of feature types: lexical, morpho-syntactic, character, and stylistic. The best results on the parliamentary transcripts dataset were achieved with the Support Vector Machine method and morpho-syntactic features; on the forum posts dataset – with the Support Vector Machine method and character features. The previous AP research on the Lithuanian language was done with parliamentary transcripts focusing on the age, gender, and political attitude dimensions (Kapočiūtė-Dzikienė et al., 2014). The best results on the age dimension were achieved with the Support Vector Machine method and a mix of lemma unigrams, bi-grams, and tri-grams; on the gender and political attitude dimensions – with the Support Vector Machine method and a mix of lemma unigrams and bi-grams.

Hence, AA and AP research using classification methods is done on parliamentary transcripts (representing normative language) and forum posts (representing non-normative language) for the Lithuanian language, but there are no reported results on literary texts so far. Since a purpose of this paper is to perform the comparative analysis with the previous research done on parliamentary transcripts and forum posts, AA and AP tasks with literary texts will be solved by keeping all experimental conditions (concerning methods and their parameters, feature types, author set sizes, dataset sizes, etc.) as similar as possible.

3 Methodology

In a straightforward form, both AA and AP problems fit a standard paradigm of a text classification problem (Sebastiani, 2002).

Thus, text documents d_i belonging to the dataset D are presented as numerical vectors capturing statistics (absolute counts in our case) of potentially relevant features. Each d_i can be attributed to one element from a closed-set of candidate authors/characteristics, defined as classes $C = \{c_j\}$.

A function φ determines a mapping how each d_i is attributed to c_j in a training dataset D^T.

Our goal is to find a method (by combining classification techniques, feature types, and feature sets) which could discover as close approximation of φ as possible.

3.1 Datasets

186 literary works (in particular, novels, novellas, essays, publicistic novels, drama) taken from the Contemporary Corpus of the Lithuanian Language (Marcinkevičienė, 2000) cover the period of 37 years from 1972 to 2012. These literary works were split into text snippets containing 2,000 symbols (including white-spaces), thus an average text document length varies from $\sim$283 to $\sim$290 tokens. Although the average text length does not fit the recommendations given by Eder (2010) (2,500 tokens for Latin and 5,000 for English, German, Polish or Hungarian) or Koppel et al. (2007) (500 tokens), these texts are not as extremely short as used in, e.g., in Luyckx (2011) or Micros and Perifanos (2011) AA research works, where reasonable results were achieved with only $\sim$60 tokens per text.

After previously described pre-processing, we composed 3 datasets:

- *LIndividual*, which was used in our AA task (see Table 1). The experiments with this dataset involved balanced/full versions and the increasing number of candidate authors (3, 5, 10, 20, 50, and 100).

- *LAge* used in our AP task by focusing on the age dimension (see Table 2) contains 6 age groups (*$\leq$29*, *30-39*, *40-49*, *50-59*, *60-69*, and *$\geq$70*).[1] The age group of any author was determined by calculating a difference between the author's birth date an the publishing date of his/her literary work. An opposite to the related research works (e.g., (Schler et al., 2006) or (Koppel et al., 2009)) we did not eliminate intermediate age groups, thus we did not simplify our task. The experiments performed with the balanced dataset versions (unless there was not enough text samples in the "main pool") of 100, 300, 500, 1,000, 2,000, 5,000 text documents in each class.

- *LGender* used in our AP task focusing on the gender dimension (see Table 2) contains 2 gender groups (*male* and *female*). The experiments performed with the balanced dataset versions of 100, 300, 500, 1,000, 2,000, 5,000 text documents in each class.

A distribution of 100 authors by their age and gender is given in Table 3. The *LAge* and *LGender* datasets contain randomly selected texts, providing no meta information about their authors.

[1] The chosen grouping is commonly used in the social studies, e.g., in the largest data archive in Europe (`http://www.gesis.org`), as well as in the Lithuanian Data Archive for Social Science and Humanities (`http://www.lidata.eu`).

Numb. of classes	Numb. of text documents	Numb. of tokens	Numb. of distinct tokens (types)	Numb. of distinct lemmas	Avg. numb of tokens in a doc.
3	450	128,622	39,306	20,099	285.83
	2,156	612,030	105,200	42,347	283.87
5	750	214,117	58,282	26,846	285.49
	3,099	877,788	136,798	51,638	283.25
10	1,500	430,849	84,838	35,424	287.23
	5,102	1,456,039	176,146	64,001	285.39
20	3,000	867,657	133,163	52,005	289.22
	8,661	2,492,637	236,505	84,566	287.80
50	7,500	217,6019	229,726	84,952	290.14
	16,317	4,721,452	343,827	124,117	289.36
100	15,000	4,347,165	332,251	120,676	289.81
	25,564	7,395,147	436,686	159,175	289.28

Table 1: Statistics about *LIndividual*: an upper value in each cell represents the balanced dataset of 150 texts in each class, a lower value– imbalanced (full). The set of authors is the same in both dataset versions.

Dataset	Numb. of text documents	Numb. of tokens	Numb. of distinct tokens (types)	Numb. of distinct lemmas	Avg. numb of tokens in a doc.
LAge	27,264	7,912,886	454,165	165,432	290.23
LGender	10,000	2,899,837	271,189	99,242	289.98

Table 2: Statistics about the balanced *LAge* and *LGender* datasets containing 5,000 text documents in each class. The *LGender* dataset is not completely balanced due to the lack of texts in the age groups of ≤*29* and ≥*70*.

	≤29	30-39	40-49	50-59	60-69	≥70
Male	5	13	13	13	13	12
Female	7	8	6	4	3	3
Total	12	21	19	17	16	15

Table 3: Distribution of authors by their age and gender.

3.2 Machine Learning Methods

In order to compare obtained AA and AP results with the previously reported, experimental conditions have to be as similar as possible. Thus, the choice of classification method was restricted to Naïve Bayes Multinomial (NBM) (introduced by Lewis and Gale (1994)) and Support Vector Machine (SVM) (introduced by Cortes and Vapnik (1995)). Both SML techniques are used in the recent AA and AP tasks due to their advantages.

3.3 Features

The choice of features (by which documents are represented) is as important as the choice of classification method. To find out what could work with the Lithuanian literary texts, we tested a big variety of different feature types, covering stylistic, character, lexical and morpho-syntactic levels:

- *sm* – style markers: an average sentence and word length; a standardized type/token ratio.
- *fwd* – function words (topic-neutral): prepositions, pronouns, conjunctions, particles, interjections, and onomatopoeias, which were automatically recognized in texts with the Lithuanian morphological analyzer-lemmatizer "Lemuoklis" (Zinkevičius, 2000).
- *chr* – (language-independent) document-level character n-grams with $n \in [2, 7]$.
- *lex* – tokens and a mix of their n-grams up to $n \in [2, 3]$ (e.g., in $n = 3$ case not only trigrams, but bi-grams and unigrams would be used as well).
- *lem* – lemmas and a mix of their n-grams up to $n \in [2, 3]$. The lemmatization was done with "Lemuoklis" which replaced recognized words with their lemmas, transformed generic words into appropriate lowercase letters and all numbers into a special tag.
- *pos* – coarse-grained part-of-speech tags (such as noun, verb, adjective, etc., determined with "Lemuoklis") and a mix of their n-grams up to $n \in [2, 3]$.
- *lexpos*, *lempos*, *lexmorf*, *lemmorf* – the compound features of *lex+pos*, *lem+pos*,

lex+morf, *lem+morf*, respectively, and a mix of their n-grams up to $n \in [2, 3]$. Here *morf* indicates a fine-grained part-of-speech tag composed of coarse-grained tag with the additional morphological information as case, gender, tense, etc.

4 Experimental Setup and Results

All experiments were carried out with the stratified 10-fold cross-validation and evaluated using the accuracy and f-score metrics.[2]

For each dataset version (described in Section 3.1) the random $\sum P^2(c_j)$ and majority $\max P(c_j)$ baselines were calculated (where $P(c_j)$ is the probability of class c_j) and the higher one of these values is presented in the following figures. The statistical significance between different results was evaluated using McNemar's (1947) test with one degree of freedom.

In all experiments we used WEKA 3.7 machine learning toolkit (Hall et al., 2009); 1,000 the most relevant features (using the types described in Section 3.3), ranked by the calculated χ^2 values; the SVM method with the SMO polynomial kernel (Platt, 1998) (because it gave the highest accuracy in the comparative experiments, done with parliamentary transcripts and forum data (Kapočiūtė-Dzikienė et al., 2015)) and the NBM method (described in Section 3.2). Remaining parameters were set to their default values.

The highest achieved accuracies (in terms of all explored feature types) with both classification techniques for AA and AP tasks are presented in Figure 1 and Figure 2, respectively. For the accuracies obtained with different feature types using the most accurate classification method and the datasets presented in Table 1, Table 2 see Table 4.

5 Discussion

All obtained results are reasonable, as they exceed the random and majority baselines.

If we compare the results in Figure 1 with the previously reported results on parliamentary transcripts and forum posts (Kapočiūtė-Dzikienė et al., 2015), SVM is not the best technique in all cases here. Despite it slightly outperforms NBM on the smaller datasets (with <20 candidate authors), but under-performs on the larger ones. We suppose that the simple NBM technique coped effectively with our AA task due to the following reasons: used literary texts are more homogeneous (a literary work/author rate is 1.86), longer ($\sim$1.34 and $\sim$6.88 times longer compared to parliamentary transcripts and forum posts, respectively), have more stable vocabulary, and clearer synonymy compared to parliamentary transcripts (covering a period of 23 years) or forum posts (covering a bunch of different topics). Moreover, the writing style of each author in literary works is expressed more clearly, therefore the drop in the accuracy when adding new authors to the dataset was not as steep as with parliamentary transcripts or forum posts. Even with 100 candidate authors the accuracy on literary texts almost reaches the threshold of 90% (see Figure 1) exceeding the results of parliamentary transcripts and forum posts by $\sim$18.6% and $\sim$54.6%, respectively. Besides, the dataset balancing boosted the accuracy on parliamentary transcripts and reduced on forum posts, but gave no noticeable impact on literary texts. Since literary texts written by the same author are very similar in style, new texts added to the dataset could not make any significant impact.

Zooming into the feature types in Table 4 allows us to state that lexical information dominates character on the smaller datasets (having $\leq$50 candidate authors). However, when the number of candidates is small ($\leq$20) many different features (based on character, lexical, lemma or compound lexical and morpho-syntactic information) perform equally well; with 50 – only unigrams of lemmas (sometimes complemented with part-of-speech tags) are significantly better compared to the rest types; with 100 – only character tri-grams are the best. The most surprising is the fact that the character feature type gave the best results on the largest dataset. Typically when dealing with morphologically rich languages and normative texts, morphological features are the most accurate (e.g., on Greek (Stamatatos et al., 2001) or on Hebrew (Koppel et al., 2006) texts). The Lithuanian language is not an exception, i.e., the experiments with parliamentary transcripts showed that token lemmas (or their n-grams) is the best feature type, whereas on forum posts (where the morphological tools could not be maximally helpful due to the text specifics) character features gave the highest accuracy. On the other hand, a robustness of character n-grams is not very surprising: i.e.,

[2] F-scores show the same trend as accuracy values in all our experiments, therefore we do not present them in the following figures and tables.

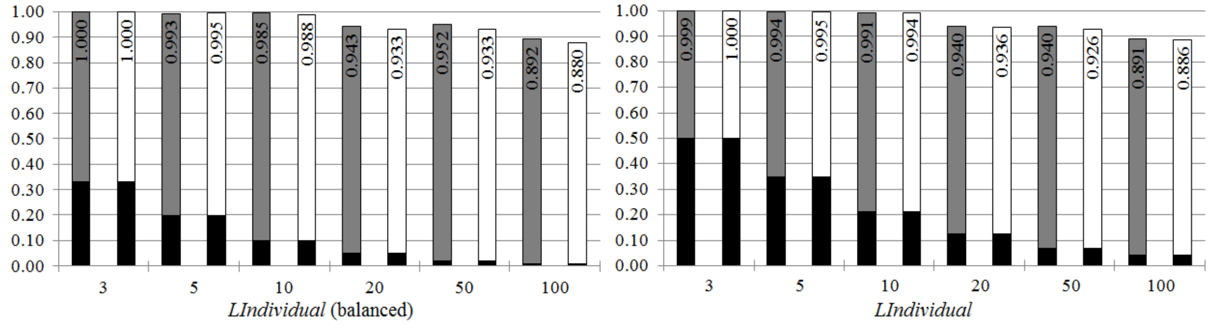

Figure 1: The accuracy (y axis) dependence on a number of candidate authors (x axis). Each column shows the maximum achieved accuracy over all explored feature types. Grey columns represent the NBM method, white – SVM, black parts represent the higher value of random/majority baselines.

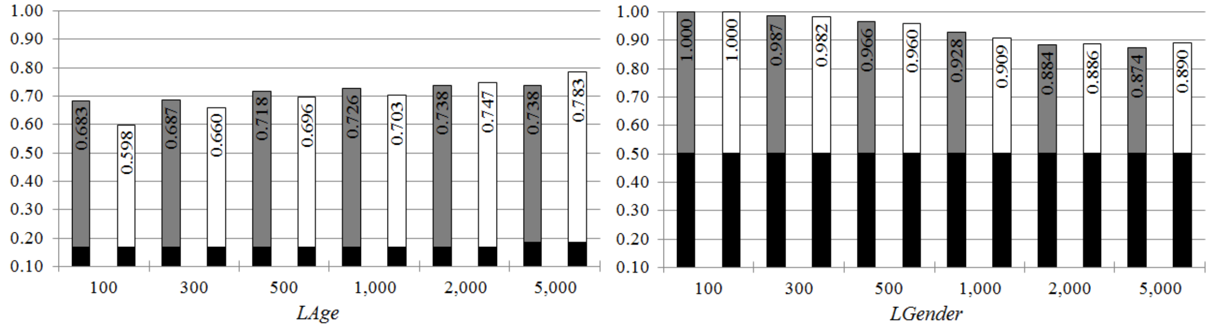

Figure 2: The accuracy (y axis) dependence on a number of instances in each class (x axis). For the other notations see the caption of Figure 1.

character n-grams can capture lexical preferences without any need of linguistic background knowledge; moreover, we used document-level character n-grams which incorporate information about contiguous words. Besides, literary texts are not too complicated for AA tasks, therefore shorter character n-grams (tri-grams in our case) are enough to capture author's style differences without mapping too obviously to specific words.

The SVM method outperformed the NBM method on the larger datasets (having more instances in each class) in all AP tasks (see the results on *LAge* and *LGender* presented in Figure 2 and on parliamentary transcripts reported by Kapočiūtė et al. (2014)). The results obtained with literary texts focusing on the age dimension do not contradict the results achieved with parliamentary transcripts: the highest boost in the accuracy is reached on the largest datasets (containing 5,000 instances in each class). The results obtained with literary texts focusing on the gender dimension are absolutely opposite, i.e., the best performance on literary texts was demonstrated with the smaller datasets, on the parliamentary transcripts – with the largest. We suppose that this unexpected situation (when the smaller datasets seem more optimal for capturing the gender characteristics) happened when instances were randomly selected from the "main pool", i.e., if the first selected instances were the most typical for the writing style of males and females, less characteristic instances added afterwards could only degrade AP performance. However, the precise answer to this question is possible only after a detailed error analysis which is planned in our future research. Nevertheless the dataset of 100, 300 or 500 instances in each class is too small to be recommended for any AP tasks.

Zooming into Table 4 allows us to state that token lemmas is the best feature type on *LAge*. Besides, lemma information (in particular, a mix of lemma tri-grams, bi-grams, and unigrams) gave the best results on parliamentary transcripts as well. Marginally the best feature type dealing with the largest *LGender* dataset is token lemmas complemented with the part-of-speech information; with the smaller *LGender* datasets various lexical, morpho-syntactic, and character feature types demonstrated high and very similar per-

Feature	***LIndividual* (balanced)**						***LIndividual***						***LAge***	***LGender***
type	**3**	**5**	**10**	**20**	**50**	**100**	**3**	**5**	**10**	**20**	**50**	**100**		
sm	0.616	0.380	0.249	0.163	0.089	0.045	0.635	0.441	0.312	0.195	0.114	0.072	0.245	0.560
fwd	0.942	0.825	0.823	0.701	0.575	0.442	0.966	0.874	0.862	0.751	0.634	0.480	0.445	0.710
chr2	0.989	0.965	0.973	0.896	0.874	0.782	0.990	0.978	0.979	0.911	0.895	0.836	0.649	0.811
chr3	0.998	0.992	0.986	0.932	0.934	**0.892**	0.998	0.991	0.989	0.920	0.921	**0.891**	0.722	0.859
chr4	0.998	0.991	0.985	0.922	0.896	0.756	0.997	0.988	0.986	0.914	0.896	0.791	0.726	0.858
chr5	0.993	0.983	0.973	0.915	0.833	0.662	0.995	0.987	0.979	0.906	0.854	0.703	0.698	0.849
chr6	0.987	0.980	0.962	0.897	0.787	0.633	0.993	0.984	0.968	0.894	0.824	0.672	0.678	0.843
chr7	0.987	0.977	0.958	0.884	0.761	0.611	0.992	0.985	0.961	0.884	0.795	0.639	0.656	0.834
lex1	**1.000**	**0.993**	0.993	0.937	0.928	0.841	0.998	**0.994**	**0.991**	0.933	0.909	0.820	0.753	0.884
lex2	**1.000**	0.991	0.991	0.933	0.916	0.840	0.998	**0.994**	0.989	0.929	0.895	0.777	0.745	0.885
lex3	**1.000**	0.992	0.991	0.934	0.916	0.841	0.998	**0.994**	0.989	0.929	0.895	0.775	0.745	0.884
lem1	0.998	0.989	0.993	**0.943**	**0.952**	0.889	**0.999**	0.990	**0.991**	**0.940**	**0.940**	0.882	**0.783**	0.889
lem2	0.998	0.991	0.992	0.936	0.941	0.886	**0.999**	0.991	0.990	0.936	0.925	0.837	0.771	0.884
lem3	0.998	0.988	0.991	0.938	0.941	0.885	**0.999**	0.989	0.990	0.937	0.925	0.834	0.769	0.884
pos1	0.702	0.632	0.549	0.356	0.218	0.143	0.890	0.641	0.553	0.368	0.209	0.142	0.340	0.616
pos2	0.882	0.768	0.784	0.612	0.516	0.408	0.891	0.780	0.789	0.641	0.529	0.430	0.437	0.649
pos3	0.909	0.797	0.817	0.691	0.615	0.516	0.909	0.801	0.818	0.689	0.618	0.533	0.441	0.648
lexpos1	**1.000**	0.992	0.993	0.940	0.925	0.838	0.998	**0.994**	0.990	0.933	0.908	0.814	0.750	0.883
lexpos2	**1.000**	0.991	0.991	0.934	0.910	0.839	0.998	0.993	0.990	0.927	0.892	0.776	0.741	0.880
lexpos3	**1.000**	0.991	0.990	0.936	0.911	0.837	0.998	0.993	0.989	0.927	0.892	0.774	0.741	0.878
lempos1	0.998	0.989	0.993	0.943	0.951	0.888	**0.999**	0.990	**0.991**	**0.940**	0.938	0.880	0.741	**0.890**
lempos2	**1.000**	0.992	0.991	0.938	0.939	0.887	0.998	0.992	0.989	0.935	0.924	0.840	0.770	0.885
lempos3	**1.000**	0.989	0.991	0.937	0.939	0.886	0.998	0.992	0.990	0.934	0.923	0.835	0.771	0.882
lexmorf1	**1.000**	0.991	**0.995**	0.939	0.926	0.838	0.998	**0.994**	0.990	0.933	0.907	0.814	0.749	0.882
lexmorf2	**1.000**	0.992	0.989	0.935	0.912	0.835	0.997	0.993	0.989	0.927	0.890	0.772	0.740	0.880
lexmorf3	**1.000**	0.991	0.990	0.935	0.911	0.835	0.997	0.992	0.989	0.927	0.890	0.771	0.739	0.879
lemmorf1	**1.000**	0.992	0.991	0.937	0.932	0.850	0.998	**0.994**	**0.991**	0.932	0.913	0.828	0.754	0.886
lemmorf2	**1.000**	0.988	0.989	0.930	0.916	0.850	0.998	0.993	0.988	0.927	0.894	0.783	0.745	0.875
lemmorf3	**1.000**	0.988	0.990	0.930	0.916	0.849	0.998	**0.994**	0.988	0.926	0.895	0.782	0.746	0.876

Table 4: The accuracy values achieved on *LIndividual* with NBM; on *LAge* and *Gender* with SVM and 5,000 instances in each class. In each column the best results are presented in bold, the results that do not significantly differ from the best one are underlined.

formance. Besides, the best feature type on parliamentary transcripts is a mix of lemma bi-grams and unigrams. However, a robustness of lemmata is not surprising having in mind that we were dealing with the morphologically complex language and normative texts.

6 Conclusions

In this paper we report the first authorship attribution and author profiling results obtained on the Lithuanian literary texts. The results are compared with previously reported on parliamentary transcripts and forum posts.

When solving the authorship attribution task we experimentally investigated the effect of the author set size by gradually increasing the number of candidate authors up to 100. The best results dealing with the maximum author set were achieved with the Naïve Bayes Multinomial method and character tri-grams. The results exceeded the baselines by ∼88.2% and reached 89.2% of the accuracy.

When solving the author profiling task we experimentally investigated the effect of balanced dataset size by gradually increasing the number of instances in each class up to 5,000. The best results for the age dimension were achieved with the maximum dataset, token lemmas, and the Support Vector Machine method; for the gender dimension very good performance was demonstrated already with the small datasets, using lemma unigrams and the Support Vector Machine method. The results focusing on the age and gender dimensions exceeded baselines by ∼60% and ∼50% reaching 78.3% and 100% of the accuracy, respectively.

The comparative analysis show that it is much easier to capture age, gender and individual author differences with literary texts than with parliamentary transcripts or forum posts.

In the future research we are planning to make the detailed error analysis, which could help us to improve the accuracy; to expand the number of authors and profiling dimensions.

Acknowledgments

This research was funded by a grant (No. LIT-8-69) from the Research Council of Lithuania.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writerprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29.

Corina Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Marco Cristani, Giorgio Roffo, Cristina Segalin, Loris Bazzani, Alessandro Vinciarelli, and Vittorio Murino. 2012. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1121–1124.

Olivier de Vel, Alison M. Anderson, Malcolm W. Corney, and George M. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4):55–64.

Maciej Eder. 2010. Does size matter? Authorship attribution, small samples, big problem. In *Digital Humanities 2010: Conference Abstracts*, pages 132–135.

Maciej Eder. 2011. Style-markers in authorship attribution a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6(1):99–114.

Mark Hall, Eibe Frank, Holmes Geoffrey, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Giacomo Inches, Morgan Harvey, and Fabio Crestani. 2013. Finding participants in a chat: authorship attribution for conversational documents. In *International Conference on Social Computing*, pages 272–279.

Matthew L. Jockers and Daniela M. Witten. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223.

Patrick Juola. 2007. Future trends in authorship attribution. In *Advances in Digital Forensics III - IFIP International Conference on Digital Forensics*, volume 242, pages 119–132.

Olga Kanishcheva. 2014. Using of the statistical method for authorship attribution of the text. In *Proceedings of the 1st International Electronic Conference on Entropy and Its Applications*, volume 1.

Jurgita Kapočiūtė-Dzikienė, Ligita Šarkutė, and Andrius Utka. 2014. Automatic author profiling of Lithuanian parliamentary speeches: exploring the influence of features and dataset sizes. In *Human Language Technologies – The Baltic Perspective*, pages 99–106.

Jurgita Kapočiūtė-Dzikienė, Ligita Šarkutė, and Andrius Utka. 2015. The effect of author set size in authorship attribution for Lithuanian. In *NODALIDA 2015: 20th Nordic Conference of Computational Linguistics*, pages 87–96.

Moshe Koppel, Dror Mughaz, and Navot Akiva. 2006. New methods for attribution of rabbinic literature. *A Journal for Hebrew Descriptive, Computational and Applied Linguistics*, 57:5–18.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Olga Vladimirovna Kukushkina, Anatoly Anatol'evich Polikarpov, and Dmitrij Viktorovich Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22Nd International Conference on Computational Linguistics*, volume 1, pages 513–520.

Kim Luyckx. 2011. Authorship attribution of e-mail as a multi-class task. In *Notebook for PAN at CLEF 2011. Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*.

Rūta Marcinkevičienė. 2000. Tekstynų lingvistika (teorija ir paktika) [Corpus linguistics (theory and practice)]. *Darbai ir dienos*, 24:7–63. In Lithuanian.

Quinn Michael McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–246.

George K. Mikros and Kostas Perifanos. 2011. Authorship identification in large email collections: experiments using features that belong to different linguistic levels. In *Notebook for PAN at CLEF 2011. Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop).*

Frederik Mosteller and David L. Wallace. 1963. Inference in an authorship problem. *Journal Of The American Statistical Association*, 58(302):275–309.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 300–314.

Jamal Abdul Nasir, Nico Görnitz, and Ulf Brefeld. 2014. An off-the-shelf approach to authorship attribution. *The 25th International Conference on Computational Linguistics*, pages 895–904.

John C. Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods – Support Vector Learning*, pages 185–208.

Tieyun Qian, Bing Liu, Li Chen, and Zhiyong Peng. 2014. Tri-training for authorship attribution with limited training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 345–351.

Tomislav Reicher, Ivan Krišto, Igor Belša, and Artur Šilić. 2010. Automatic authorship attribution for texts in Croatian language using combinations of features. In *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6277, pages 21–30.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micromessages. In *Empirical Methods in Natural Langauge Processing*, pages 1880–1891.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Thamar Solorio, Sangita Pillay, Sindhu Raghavan, and Manuel Montes-y Gómez. 2011. Modality specific meta features for authorship attribution in web forum posts. In *The 5th International Joint Conference on Natural Language Processing*, pages 156–164.

Rui Sousa-Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio C. Oliveira, and Belinda Maia. 2011. 'twazn me!!! ;(' Automatic authorship analysis of micro-blogging messages. In *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems*, pages 161–168.

Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556.

Urszula Stańczyk and Krzysztof A. Cyran. 2007. Machine learning approach to authorship attribution of literary texts. *International Journal of Applied Mathematics and Informatics*, 1(4):151–158.

Hans Van Halteren, R. Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12:65–77.

Andelka Zečević and Miloš Utvić. 2012. An authorship attribution for Serbian. In *Local Proceedings of the Fifth Balkan Conference in Informatics*, pages 109–112.

Andelka Zečević. 2011. N-gram based text classification according to authorship. In *Proceedings of the Student Research Workshop associated with Recent Advances in Natural Language Processing*, pages 145–149.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of the Second AIRS Asian Information Retrieval Symposium*, pages 174–189.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.

Vytautas Zinkevičius. 2000. Lemuoklis – morfologinei analizei [Morphological analysis with Lemuoklis]. *Darbai ir dienos*, 24:246–273. In Lithuanian.

Ana Zwitter Vitez. 2012. Authorship attribution: specifics for Slovene. *Slavia Centralis*, 1(14):75–85.