The Effect of Author Set Size in Authorship Attribution for Lithuanian

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

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

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

Abstract

This paper reports the first authorship attribution results based on the effect of the author set size using automatic computational methods for the Lithuanian language. The aim is to determine how fast authorship attribution results are deteriorating while the number of candidate authors is gradually increasing: i.e. starting from 3, going up to 5, 10, 20, 50, and 100. Using supervised machine learning techniques we also investigated the influence of different features (lexical, character, morphological, etc.) and language types (normative parliamentary speeches and non-normative forum posts).

The experiments revealed that the effectiveness of the method and feature types depends more on the language type rather than on the number of candidate authors. The content features based on word lemmas are the most useful type for the normative texts, due to the fact that Lithuanian is a highly inflective, morphologically and vocabulary rich language. The character features are the most accurate type for forum posts, where texts are too complicated to be effectively processed with external morphological tools.

1 Introduction

Authorship Attribution (AA) is the task of identifying who, from a set of candidate authors, is an actual author of a given anonymous text document. This prediction is based on a human "stylometric fingerprint" notion: i.e. a specific, individual, persistent, and uncontrolled habit to express thoughts with a unique set of linguistic means. Van Halteren (2005) has gone so far as to name this phenomenon a "human stylome" in the deliberate analogy to the DNA "genome". However, Juola (2007) argues that such strict implications may not be absolutely correct, because the "genome" is stable, but the human style tends to evolve over time. Nevertheless a "stylome" can still be added to human biometrics, next to voice, gait, keystroke dynamics, handwriting, etc.

Starting from Mendenhall (1887) AA is one of the oldest computational linguistics problems, which is especially highly topical nowadays. For a long time in the past the main AA applications were restricted to the literary texts only. But the constant influx of anonymous electronic text documents, especially on the Internet, and the popularity of automatic methods opened the gate to a number of new applications in forensic analysis and electronic commerce. In addition to literary research the practical problems from the plagiarism detection (Stamatatos, 2011), the identification of harassment and threatening (Tan et al., 2013) to tracking authors of malicious source code (Alrabaee et al., 2014) gained even greater prominence. This led to experiments with different datasets, such as e-mails (de Vel et al., 2001; Abbasi and Chen, 2008), 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 progress of the development of computational linguistic methods that are able to cope with the emerged problems.

Despite that many computational linguistic tasks can be solved accurately only relying on efforts of domain-experts, it is very time consuming, expensive, and perhaps the most limiting way for AA, moreover, which provides no explicit measure how attributions are made. The alternative way is a manually composed set of rules capable to take attribution decisions automatically. Unfortunately, rule-based systems usually are very complex, unwieldy, and thus not robust to any changes in the domain, language or author characteristics, therefore it is rather difficult to make any updates. Moreover, when dealing with hundreds (e.g. in Luyckx and Daelemans (2008), Luyckx (2010)) or thousands of candidate authors (e.g. in Koppel et al. (2011) 10,000 authors; in Narayanan et al. (2012) – 100,000) the possibility to create an effective rule set goes far beyond human potential limits. Ultimately, AA task can be solved using the machine learning (Sebastiani, 2002): i.e. by training the classifiers and later using them to predict the authorship of unseen texts. Moreover, it can be easily adjusted to new applications or domains and even generalized well to drifts in the author characteristics. Due to all these advantages, the machine learning paradigm became dominant and remained the most popular till nowadays. Therefore our focus in this paper is also on the machine learning methods.

2 Related Work

Despite rare attempts to deal with unlabeled data, e.g. Nasir et al. (2014), Qian et al. (2014), a typical AA problem fits the standard paradigm of the supervised machine learning. It means that the training dataset containing texts of known authors is available and can be used to create the model able to predict the authorship of unknown texts from the same closed-set of the candidate authors in the future. Algorithmically, it involves a variety of different methods (for the detailed review see Stamatatos (2009)) ranging from probabilistic approaches (Seroussi et al., 2011), compression models (Oliveira et al., 2013) to Vector Space Models (Stamatatos, 2008). In general all methods can be distinguished according to whether they treat each training text individually (instance-based) or cumulatively by concatenating texts written by the same author into one (profile-based). Intuitively, profile-based approaches should have advantages over instance-based when text documents are very concise, thus concatenation helps to create sufficiently long document for capturing its style; but on the other hand instance-based approaches are better suited for the sparse data scenario. Some comparative experiments on the AA after testing Decision Trees (DTs), Back Propagation Neural Networks (BPNNs) and Support Vector Machines (SVMs) revealed that SVMs and BPNNs achieved significantly better performance compared to DTs (Zheng et al., 2006). Zhao and Zobel (2005) proved that k-Nearest Neighbor (kNN) approach produces better results compared to both Naïve Bayes (NB) and DTs. Jockers and Witten (2010) report that Delta method outperforms popular SVMs. Savoy (2012) proposes new classification scheme based on the specific vocabulary and experimentally proves that it performs better than Principal Component Analysis (PCA) and slightly better than Delta approach; Savoy (2013) also shows that LDA (Latent Dirichlet Allocation) classification scheme can surpass two classical AA approaches – i.e. Delta rule and chi-squared distance. Nevertheless, the precise comparison of methods is still difficult due to the lack of suitable benchmark data. Besides, the results are affected not only by the selected classification method itself, but by preprocessing techniques, author set sizes, language characteristics, etc. However the most crucial factor is probably the selected type of features.

The first modern work in AA (different from traditional human-expert techniques) was described by Mosteller and Wallace (1963). They demonstrated promising AA results on *The Federalist papers* using Bayesian methods applied on frequencies of a small set of function words (including articles, prepositions and conjunctions) as stylistic features in the text. Since this pioneering study and until 1990s AA was based on quantitative features (so-called style markers) such as a sentence or word length, syllables per word, type-token ratio, vocabulary richness functions, lexical repetition, etc. In fact all these stylometric features are considered to be suitable only for homogeneous long texts (>1,000 words) and for datasets where the number of candidate authors is limited. Lately other feature types– in particular, lexical, syntactic, semantic, or character –treating texts as the sequence of tokens or characters became more popular. A huge number of these features have been presented so far, but we will focus only on the most popular and the most accurate ones. The most common example of the lexical feature type is a simple bag-of-words representation which is considered to be topic-dependent therefore should be avoided when the distribution over authors coincides the distribution over different topics (not to solve topic-classification problem instead of AA). Besides token n-grams are also considered to capture content-specific instead of stylistic information. The most popular topic-neutral lexical solution, carrying no semantic information, is the function words (articles, conjunctions, prepositions, pronouns, etc.). Various authors use different lists of function words, varying from 150 (Abbasi and Chen, 2005) to 675 (Argamon et al., 2007) words, but providing very little information about how these lists were composed. The effectiveness of syntactic and semantic features usually rely on the accuracy of external linguistic tools (e.g. part-of-speech taggers, parsers) or exhaustiveness of additional data resources (e.g. thesauruses or databases). Although used alone they hardly can outperform lexical features, but often improve the results used in the combination (Gamon, 2004). However, character features (character n-grams, in particular) are considered the most important document representation type in authors' style detection: they are topic-neutral, language-independent, able to capture style through lexical and contextual information, and are tolerant to grammatical errors. Application-specific features are highly dependent on the solvable problem, e.g. positions of hashtags, smileys, punctuation are important style detectors in tweets (Sousa-Silva et al., 2011).

The majority of surveyed research works deal with Germanic languages, providing no guidance what could work the best with morphologically rich, highly inflective, derivationally complex, and relatively free word order languages such as Lithuanian. Starting from 1971 (Pikčilingis, 1971) lots of descriptive linguistic works are done on the AA for the Lithuanian language (the review in Žalkauskaitė (2012)). Besides, the pioneering and as far as we know the only work using automatic methods on the Lithuanian texts is described in (Kapočiūtė-Dzikienė et al., 2014). However, their experiments have been made only with the normative Lithuanian language, few authors, and small training data; therefore findings are not robust to make the generalizations about which method is the best and which feature type is the most reliable for solving AA problem in general. Consequently in this research we will try to overcome all mentioned shortcomings by experimenting with different language types (normative and Internet forum data) and increasing number of candidate authors (up to one hundred).

3 Methodology

In essence, AA problem is a task which can be formally described as follows.

The dataset D contains text documents d_i attributed to a closed-set of candidate authors (defined as classes) $C = \{c_j\}$.

The training dataset D^T (where $D^T \subset D$) is composed of training instances: i.e. documents d_i with a known authorship c_j: $\{\langle d_i, c_j \rangle\}$.

The function φ determines the mapping (about characteristics in styles of the authors) how each d_i is attributed to c_j in D^T.

Our goal is using D^T to train a classifier and to create the model φ', which could be as close approximation of φ as possible.

3.1 The Datasets

All our experiments were carried out on 2 datasets to make sure that findings generalize over different domains and language types:

- *ParlTranscr*[1] (see Table 1) contains unedited transcripts of parliamentary speeches and debates, thus representing formal spoken but normative Lithuanian language. All transcripts are from regular parliamentary sessions and cover the period of 7 parliamentary terms starting on March 10, 1990 and ending on December 23, 2013. Very long (>1,000 words) and very short (<100 words) texts were removed from the dataset to avoid speeches written by non-parliamentarians, but by someone else and to avoid less informative text samples, respectively. Afterwards we selected 100 authors with the largest number of texts, but making sure that the selected candidates are distributed over different parliamentary terms (to avoid topic classification) and party groups (to avoid ideology-based classification).

- *LRytas*[2] (see Table 2) contains forum data full of informal words, foreign language insertions, word shortenings, emoticons, and diacritic eliminations, thus represents the informal non-normative Lithuanian language. The forum has 11 general topics (such as "Business", "Politics", "Sports", etc.). Very short texts (<10 words) were not included into

[1] Downloaded from http://www3.lrs.lt/pls/inter/w5_sale.

[2] Crawled on March 19, 2014 from http://forum.lrytas.lt/forum_show.pl.

the dataset. Afterwards we selected 100 authors having the largest number of texts, but making sure that selected candidates would be distributed over different topics (to avoid topic classification).

3.2 Classification

In this paper we focus on the supervised machine learning techniques (Kotsiantis, 2007) applied to the text categorization (Sebastiani, 2002) and used for the AA (Stamatatos, 2009).

The aim of our task is to find a method, which could distinguish the distinct authors from each other by creating a model for the best approximation of the authors' style. For this reason we explored two supervised machine learning approaches:

- *Support Vector Machine* (SVM) (introduced by Cortes and Vapnik (1995)) is a discriminative instance-based approach, which is currently the most popular text classification technique, efficiently handling a high dimensional feature spaces (e.g. maximum ∼295 thousand features in the imbalanced 100 authors *ParlTrascr* dataset, ∼84,4 thousand in *LRytas*); sparseness of the feature vectors (only ∼215 non-zero feature values among ∼295 thousand in *ParlTranscr* and ∼42 among ∼84,4 thousand in *LRytas*); and does not perform aggressive feature selection, which may result in a loss of information and degrade the accuracy (Joachims, 1998).

- *Naïve Bayes Multinomial* (NBM) (introduced by Lewis and Gale (1994)) is a generative profile-based approach, which is often selected due to its simplicity: Naïve Bayes assumption about the feature independence allows parameters of each feature to be learned separately; the method performs especially well when the number of features having equal significance is large; it is very fast and does not require huge data storage resources; besides, this Bayesian method is often selected as the baseline approach.

However, it is important to notice that the choice of classification algorithm is not more important than the choice of feature types by which texts have to be represented.

3.3 Feature Extraction

In our research we explored the impact of the most popular or/and accurate individual and compound feature types, covering stylistic, character, lexical, and morpho-syntactic levels:

- *usm* – ultimate style markers: average sentence and word length in a text document; standardized type/token ratio (STTR). Although we assume that this archaic stylometric feature type will definitely give very poor classification results, it still has to be tested for comparison reasons.

- *chrN* – document-level character n-grams: context-free character feature type (where *N* = [2;7] in our experiments). It considers successions of *N* characters including spaces and punctuation marks, e.g., *chr7* of phrase "authorship attribution" produces the following character n-grams: "authors", "uthorsh", "thorshi", "horship", "orship_", "rship_a", etc.[3] By many researchers this feature type was proved to be one of the best (or even the best) to tackle AA problems.

- *fwd* – function words: the content-free lexical feature type which includes prepositions, pronouns, conjunctions, particles, interjections, and onomatopoeias. Instead of relying on the pre-established and stable list of the function words, we identified them by applying the Lithuanian morphological analyzer-lemmatizer "Lemuoklis" (Zinkevičius, 2000; Daudaravičius et al., 2007). This feature type by consensus is considered as the topic-neutral and was proved to be a relatively good identifier of the writing style by many researchers.

- *lexN* – token n-grams: the most popular content-specific lexical feature type which involves a bag-of-words (*N* = 1) or interpolation of token n-grams (*N* = [2;3] in our experiments), e.g., *lex1* of the phrase "authorship attribution problem" produces 3 bag-of-words: "authorship", "attribution", and "problem"; *lex2*: 3 bag-of-words plus token bigrams "authorship attribution", and "attribution problem"; *lex3*: 3 bag-of-words, 2 to-

[3] This and the following examples will be given in English instead of Lithuanian for the clarity reasons.

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	600	156,107	21,439	8,608	260.18
	16,804	3,457,093	107,950	35,525	205.73
5	1,000	239,288	27,983	10,864	239.29
	22,476	4,585,493	132,623	42,620	204.02
10	2,000	451,638	38,952	14,076	225.82
	34,307	6,821,083	157,409	49,470	198.82
20	4,000	927,411	63,456	21,310	231.85
	50,532	10,254,271	204,043	61,443	202.93
50	10,000	2,475,615	107,029	33,308	247.56
	77,005	16,478,475	254,966	75,563	213.99
100	20,000	4,728,411	151,836	45,441	236.42
	98,999	21,295,515	295,046	86,770	215.11

Table 1: Composition of *ParlTranscr*: the upper value in each cell represents a balanced dataset (200 instances in each class), the lower – imbalanced (full). The set of authors is identical in the both datasets.

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	30	1,252	792	615	41.73
	3,567	137,768	30,830	16,726	38.62
5	50	1,722	1,049	781	34.44
	4,579	166,512	36,267	19,271	36.36
10	100	3,913	2,191	1,572	39.13
	6,209	244,947	49,648	26,603	39.45
20	200	8,876	4,287	2,910	44.38
	8,470	351,285	63,363	33,377	41.47
50	500	21,942	8,980	5,725	43.88
	11,155	468,466	76,861	40,057	42.00
100	1,000	44,375	15,290	9,443	44.38
	12,888	545,405	84,482	44,211	42.32

Table 2: Composition of *LRytas*: the upper value in each cell represents a balanced dataset (10 instances in each class), the lower – imbalanced (full). The set of authors is identical in the both datasets.

ken n-grams plus one trigram “authorship attribution problem”.

- *lemN* – n-grams of token lemmas: the content-specific lexical feature type which involves lemmas based on the word tokens (N = 1) or their interpolation (N = [2;3] in our experiments). “Lemuoklis” replaces words with their lemmas, transforms recognized generic words into the lower-case and replaces all numbers with a special tag. We assume that this feature type should reduce the number of types significantly (especially for *ParlTranscr*) which should result in creation of more robust models and higher classification accuracy.

- *posN* – n-grams of part-of-speech tags: the content-free morpho-syntactic feature type which involves coarse-grained part-of-speech tags based on word tokens (N = 1) or their interpolation (N = [2;3] in our experiments). Coarse-grained part-of-speech tags (such as noun, verb, adjective, etc.) are also determined by “Lemuoklis”.

- *lexposN*, *lemposN*, *lexmorfN*, *lemmorfN* – the aggregated features which involve unigrams (N = 1) of concatenated features or their interpolation (N = [2;3] in our experiments): *lex&pos*, *lem&pos*, *lex&morf*,

lex&morf, respectively, where *morf* indicates the string of the concatenated fine-grained morphological values for case, gender, tense, mood, etc., determined by "Lemuoklis", e.g., *lexpos2* of phrase "interesting problem" produces two unigrams "interesting_ADJ", "problem_NOUN" plus one bigram "interesting_ADJ problem_NOUN".

4 Experimental Set-Up and Results

Our aim is to explore different classification methods (see Section 3.2), feature types (see Section 3.3) and to answer the main questions:

- How the author set size affects results, when having 3, 5, 10, 20, 50, and 100 candidate authors? The candidate author selection is done depending on the number of their texts: the authors with the most texts are selected first.
- How the language type influences results, when *ParlTranscr* contains texts of normative, but *LRytas* of non-normative language?

All experiments were carried out with the stratified 10-fold cross-validation and evaluated using accuracy and micro/macro average F-score metrics. Since F-scores showed the same accuracy trend in all our experiments, we do not present them in the following figures and tables. For each dataset random ($\sum P^2(c_j)$) and majority ($\max P(c_j)$) baselines (where $P(c_j)$ is the probability of class c_j) were calculated, but only the higher values were presented in the following figures. In order to determine whether the differences between obtained values are statistically significant we performed McNemar's (McNemar, 1947) test with one degree of freedom.

In our experiments we used chi-squared feature extraction method, SMO polynomial kernel (because it gave the highest accuracy in our preliminary control experiments) with SVM and NBM implementations in the WEKA machine learning toolkit (Hall et al., 2009), version 3.6. All remaining parameters were set to their default values.

For the effect of used method see Figure 1 and feature type see Table 3 and Table 4.

5 Discussion

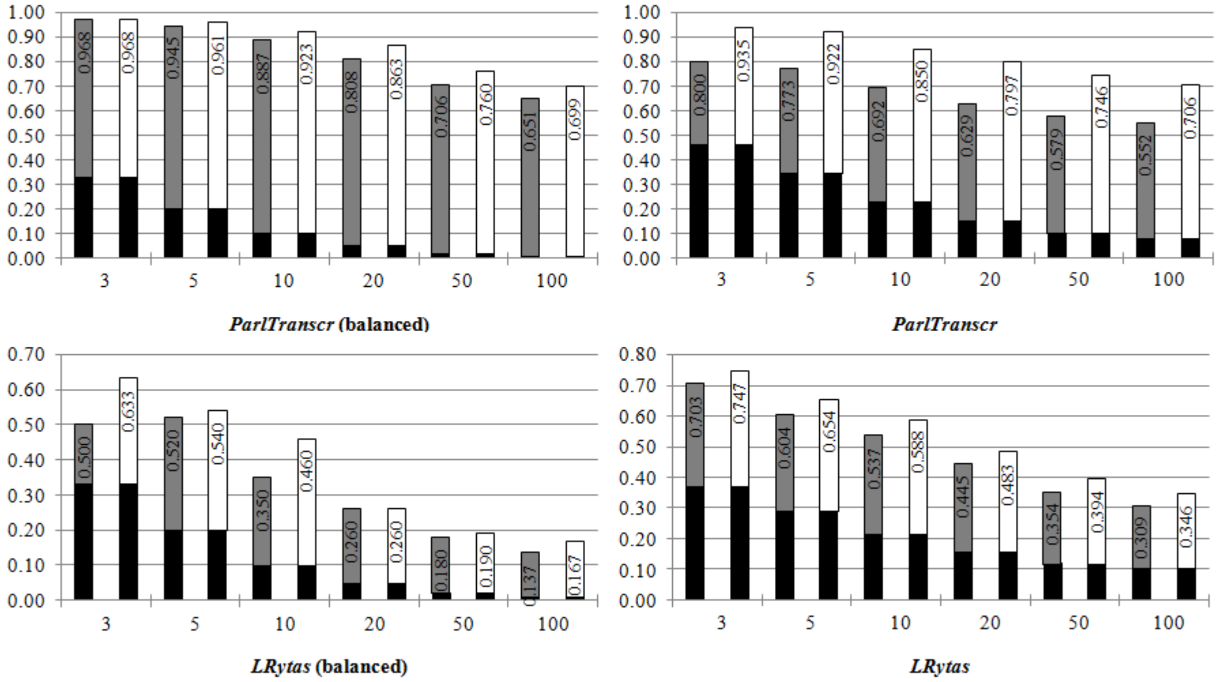
Zooming into the results presented in Figure 1, allows us to report the following statements:

All obtained results are reasonable and appropriate for our solving task, because they exceed random and majority baselines. However, SVM is a much better selection, as it always outperformed NBM, except for a couple of cases when the both methods achieved the same accuracy.

If compared the same number of candidate authors, the accuracy of *LRytas* is always much lower compared to *ParlTranscr*. This could be due to the language type, text length, and training dataset size. The comprehensive expert analysis revealed that parliamentarians use official language with the larger but more steady dictionary. Moreover, their speeches or debates are carefully transcribed, thus there are no grammatical errors and diacritic eliminations. Whereas in *LRytas* different forum texts posted by even the same author are written in different manners, thus the quality of texts varies (sometimes more typing errors or abbreviations). Since the non-normative language is always much harder to deal with, the accuracy is lower. The second reason is the length of classified texts: it is always easier to predict the author from longer text samples. As we can see from the Table 1 and Table 2 the texts in *ParlTranscr* are more than 5 times longer compared to *LRytas* texts. Besides, in our experiments we were using 10-fold cross-validation, thus having 9/10 of all text samples for training, e.g., when dealing with the imbalanced datasets and 100 candidate authors, *ParlTransc* has 7 times more text documents and 3 times more different tokens (types) compared to *LRytas* (see Table 1 and Table 2). Consequently, the bigger variety in the training data helps to create more comprehensive models which in turn are more robust in the classification stage.

When increasing the number of candidate authors, we are also making the task more difficult, thus the accuracy is gradually dropping. However the decline is much steeper for *LRytas* compared to *ParlTranscr*, e.g. the increase from 3 to 100 candidate authors using SVM for balanced and imbalanced *ParlTranscr* produces the decrease of 26.9% and 22.9%, respectively; while for *LRytas* it is 46.6% and 40.1%. Having more candidate authors the task becomes more difficult, therefore all previously mentioned problems (language type, text length, training dataset size), become even more detriment.

The balancing decreases training data, thus negatively affects AA results for *LRytas* and SVM's results with 100 authors for *ParlTranscr* (this confirms a statement in Manning and Schütze (1999)),



***ParlTranscr* (balanced)**

ParlTranscr

***LRytas* (balanced)**

LRytas

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

but has opposite effect on *ParlTranscr*. This might happened due to a successful random selection of instances for the balanced dataset. In the imbalanced experiment the major 3 authors already has the texts which are not appropriate to express their style as good as in the balanced dataset, thus a negative influence not only persists, but may increase when adding more authors to the dataset.

In our experiments we tested all the most popular currently known feature types (29 in total) used for AA. Zooming into the results reported in Table 3 and Table 4 allows us to make the following statements.

When analyzing the normative Lithuanian language (as it is in *ParlTranscr*) the content information is very important for achieving high classification accuracy. Moreover, the feature type based on word lemmas is marginally the best in 8 of 12 times (with 5, 10, 20, 50 and 100 candidate authors with balanced and 10, 50 and 100 authors with imbalanced dataset). When having 5 or 20 authors with imbalanced dataset a bit longer patterns $N = 2$ and $N = 3$, respectively, of word lemmas give the best results. Despite that the part-of-speech information when used alone is definitely not the best selection, but in concatenation with lemmas (when $N = 2$) it can boost the performance and become the best feature type with 3 authors. Considering information about statistical significance between different results, and ignoring small variations depending on the number of authors, we can state that in general the best feature type for *ParlTranscr* dataset is based on the lemma and part-of-speech information. It is not surprising due to the fact that we were dealing with the Lithuanian language which is highly inflective, morphologically and vocabulary rich; moreover we were dealing with the normative language; therefore morphological tools were maximally helpful for this dataset.

When dealing with forum posts in *LRytas*, the picture is absolutely different. Marginally the best feature type in most of the cases is not based on the content information, thus, it is not based on the lemma information. Document-level character bigrams give the best results in 9 of 12 cases with the small exceptions (100 candidate authors with balanced and 3 and 5 authors with imbalanced datasets), where the credit is given to the content lemma information again. It is not surprising, since we were dealing with the non-normative Lithuanian language texts full of errors, diacritic eliminations, and words out of the standard Lithuanian language dictionary; moreover,

Feature type	ParlTranscr (balanced)						ParlTranscr					
	3	5	10	20	50	100	3	5	10	20	50	100
usm	0.488	0.430	0.272	0.155	0.069	0.037	0.581	0.435	0.299	0.202	0.133	0.103
fwd	0.792	0.763	0.662	0.518	0.392	0.324	0.801	0.753	0.642	0.555	0.461	0.398
chr3	0.938	0.945	0.901	0.819	0.699	0.627	0.904	0.890	0.804	0.747	0.680	0.633
chr4	0.938	0.945	0.893	0.813	0.685	0.601	0.906	0.887	0.802	0.743	0.674	0.625
lex1	0.953	0.936	0.900	0.816	0.708	0.635	0.927	0.911	0.832	0.774	0.706	0.659
lem1	0.960	**0.961**	**0.922**	**0.862**	**0.760**	**0.699**	0.931	0.920	**0.850**	0.796	**0.746**	**0.706**
lem2	0.962	0.958	0.910	0.852	0.753	0.691	0.932	**0.922**	0.847	0.797	0.740	0.702
lem3	0.957	0.954	0.914	0.849	0.753	0.690	0.933	0.921	0.847	**0.797**	0.737	0.701
pos3	0.742	0.715	0.643	0.509	0.359	0.261	0.807	0.755	0.655	0.558	0.439	0.364
lexpos1	0.962	0.943	0.906	0.815	0.705	0.637	0.926	0.912	0.835	0.774	0.708	0.659
lempos1	0.960	0.956	0.918	0.851	0.750	0.690	0.934	0.921	0.847	0.795	0.742	0.701
lempos2	**0.968**	0.954	0.913	0.841	0.741	0.682	**0.935**	0.919	0.846	0.795	0.738	0.698
lexmorf1	0.953	0.941	0.900	0.812	0.703	0.632	0.922	0.911	0.831	0.771	0.705	0.657
lemmorf1	0.958	0.936	0.907	0.822	0.708	0.646	0.925	0.913	0.835	0.778	0.715	0.671

Table 3: Accuracy values with SVM and various feature types for *ParlTrascr* dataset. Only the best results in terms of *N* of each feature type are reported. The best results for different author set sizes (in columns) are in bold; results that do not statistically significant differ from the best result are underlined.

Feature type	LRytas (balanced)						LRytas					
	3	5	10	20	50	100	3	5	10	20	50	100
usm	0.267	0.360	0.250	0.085	0.038	0.027	0.443	0.343	0.251	0.183	0.139	0.121
fwd	0.300	0.280	0.170	0.105	0.078	0.045	0.587	0.459	0.375	0.293	0.225	0.197
chr2	**0.500**	**0.520**	**0.350**	**0.260**	**0.180**	0.135	0.698	0.584	**0.537**	**0.445**	**0.354**	**0.309**
lex1	0.467	0.400	0.320	0.175	0.136	0.103	0.695	0.578	0.512	0.397	0.323	0.281
lem1	0.433	0.380	0.310	0.165	0.172	0.128	0.696	**0.604**	0.525	0.418	0.336	0.230
lem2	0.400	0.280	0.210	0.205	0.152	0.127	0.687	0.583	0.506	0.409	0.328	0.287
pos1	0.400	0.340	0.220	0.180	0.106	0.060	0.609	0.486	0.407	0.304	0.233	0.202
pos2	0.367	0.280	0.260	0.150	0.082	0.063	0.649	0.536	0.469	0.353	0.275	0.238
lexpos1	0.467	0.440	0.290	0.225	0.128	0.097	0.689	0.570	0.500	0.394	0.316	0.281
lempos1	0.467	0.380	0.280	0.140	0.144	**0.137**	0.692	0.592	0.526	0.413	0.337	0.298
lempos2	0.367	0.260	0.200	0.190	0.146	0.118	0.697	0.586	0.508	0.407	0.327	0.289
lexmorf1	0.400	0.400	0.240	0.220	0.124	0.087	0.695	0.571	0.501	0.396	0.315	0.276
lemmorf1	0.367	0.340	0.240	0.140	0.130	0.105	**0.703**	0.570	0.506	0.395	0.318	0.278

Table 4: Accuracy values with SVM and various feature types for *LRytas* dataset. For other notations see the caption of Table 3.

even in forums for the registered users the identity of the author is not 100% certain. Despite all these findings about character n-grams, we cannot strongly state that it is the very best feature type for our non-normative texts, because the differences between other content-based feature types are not always statistically significant.

6 Conclusions and Future Work

In this paper we report the first authorship attribution results based on the exploration of the effect of the author set size when dealing with normative and non-normative Lithuanian language texts and using supervised machine learning techniques.

We experimentally have determined that the effect of feature types depend more on the language type used in the dataset than on the number of candidate authors. Using parliamentary data (thus normative Lithuanian language) the best feature types are based on the morpho-syntactic information generated by the external grammatical tools. The results exceed baseline by 62.7% and reach even 70.6% of accuracy with 100 of candidate authors. Using forum posts (thus non-normative texts) the best feature types are however based on the character n-grams. The results exceed baseline by 20.7% and reach 30.9% of accuracy.

In the future research we are planning to further expand the number of candidate authors up to several thousands or even tens of thousands; to experiment more with non-normative Lithuanian language (blog data, tweets, etc.) and to reveal if the same statements about feature types and methods are still valid.

Acknowledgments

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

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems*, 20(5):67–75.

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.

Saed Alrabaee, Noman Saleem, Stere Preda, Lingyu Wang, and Mourad Debbabi. 2014. OBA2: An Onion approach to Binary code Authorship Attribution. *Digital Investigation*, 11(1):S94–S103.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic Text Classification Using Functional Lexical Features: Research Articles. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

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.

Vidas Daudaravičius, Erika Rimkutė, and Andrius Utka. 2007. Morphological annotation of the Lithuanian corpus. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 94–99.

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.

Michael Gamon. 2004. Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617.

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.

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.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with many Relevant Features. In *10th European Conference on Machine Learning*, volume 1398, pages 137–142.

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 – The International Federation for Information Processing*, volume 242, pages 119–132.

Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. 2014. Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches. In *17th International Conference on Text, Speech, and Dialogue*, pages 93–100.

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

Sotiris B. Kotsiantis. 2007. Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24.

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. 2010. *Scalability Issues in Authorship Attribution*. Ph.D. thesis, University of Antwerp, Belgium.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

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*, 9:237–246.

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.

Walter Ribeiro Oliveira, Edson Justino, and Luiz S. Oliveira. 2013. Comparing compression models for authorship attribution. *Forensic Science International*, 228(1-3):100–104.

Juozas Pikčilingis. 1971. *Kas yra stilius?[What is the style?].* Vaga, Vilnius, Lithuania. (in Lithuanian).

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.

Jacques Savoy. 2012. Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics*, 19(2):132–161.

Jacques Savoy. 2013. Authorship Attribution Based on a Probabilistic Topic Model. *Information Processing and Management*, 49(1):341–354.

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

Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2011. Authorship Attribution with Latent Dirichlet Allocation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 181–189.

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. 2008. Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Information Processing and Management*, 44(2):790–799.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2011. Plagiarism Detection Using Stopword N-Grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527.

Enhua Tan, Lei Guo, Songqing Chen, Xiaodong Zhang, and Yihong Zhao. 2013. UNIK: Unsupervised Social Network Spam Detection. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, pages 479–488.

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(1):65–77.

Gintarė Žalkauskaitė. 2012. *Idiolekto požymiai elektroniniuose laiškuose. [Idiolect signs in e-mails].* Ph.D. thesis, Vilnius University, Lithuania. (in Lithuanian).

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]. In *Darbai ir Dienos*, volume 24, pages 246–273. (In Lithuanian).