Automatic Authorship Attribution for Texts in Croatian Language Using Combinations of Features

Tomislav Reicher, Ivan Krišto, Igor Belša, and Artur Šilić

Faculty of Electrical and Computing Engineering
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{tomislav.reicher, ivan.kristo, igor.belsa, artur.silic}@fer.hr

Abstract. In this work we investigate the use of various character, lexical, and syntactic level features and their combinations in automatic authorship attribution. Since the majority of text representation features are language specific, we examine their application on texts written in Croatian language. Our work differs from the similar work in at least three aspects. Firstly, we use slightly different set of features than previously proposed. Secondly, we use three different data sets and compare the same features across those data sets to draw stronger conclusions. The data sets that we use consist of articles, blogs and books written in Croatian language. Finally, we employ a classification method based on a strong classifier. We use the Support Vector Machine algorithm to learn classifiers which achieve excellent results from 91% accuracy and 91% F_1 measure up to 99% accuracy and 99% F_1 measure.

Key words: author attribution, function words, POS n-grams, feature combinations, SVM.

1 Introduction

Automatic authorship attribution is a process in the field of text classification dealing with author identification of a given text. It can be interpreted as a problem of text classification based on linguistic features specific to certain authors. The main concern in computer-based authorship attribution is defining the appropriate characterization of the text. Such characterization should capture the writing style of the authors [4].

Authorship attribution can help in document indexing, document filtering and hierarchical categorization of web pages [12]. These applications are common in the field of information retrieval. It must be noted that authorship attribution differs from plagiarism detection. Plagiarism detection attempts to detect similarities between two substantially different pieces of work. However, it is unable to determine if they were produced by the same author or not [5].

The problem of authorship attribution can be divided into three categories [21]: binary, multi-class and single-class (or one-class) classification. Binary classification solves the problem when the data set contains the texts written by

one of two authors. Multi-class classification is a generalization of the binary classification when there are more than two authors in the data set. One-class classification is applied when only some of the texts from the data set are written by a particular author while the authorship of all the other texts is unspecified. This classification ascertains whether a given text belongs to a single known author or not.

This paper presents a study of multi-class classification for the texts written in the Croatian language. The work is oriented on the combination and evaluation of different text representation features on different data sets. The rest of the paper is organized in the following manner. Section 2 discusses related work in authorship attribution and similar problems. Section 3 introduces different types of text representation features we have utilized. Section 4 describes the classification, Section 5 describes the used data sets and Section 6 presents evaluation methods and experiment results. The conclusion and future work are given in Section 7.

2 Related Work

There are several approaches to author attribution in respect of different text representation features used for the classification. Based on those features, the following taxonomy can be made [16]: *character features*, *lexical features*, *syntactic features*, *semantic features* and *application-specific features*. The following paragraphs describe character, lexical and syntactic features in more depth and relate our work with the existing research.

Character features are the simplest text representation features. They consider text as a mere sequence of characters and are thereby usable for any natural language or corpus. Various measures can be defined, such as characters frequencies, digit frequencies, uppercase and lowercase character frequencies, punctuation marks frequencies, etc. [5]. Another type of character based features, which has been proven as quite successful [14,15], considers extracting frequencies of character n-grams.

Text representation using *lexical features* is characterized by dividing the text into a sequence of tokens (words) that group into sentences. Features directly derived from that representation are the length of words, the length of sentences and vocabulary richness. This types of features have been used in [7,13] . Results achieved demonstrate that they are not sufficient for the task mostly due to their significant dependence on the text genre and length. However, taking advantage of features based on frequencies of different words, especially function words, can produce fairly better results [1,10,18,21]. Analogous to character n-grams, word n-gram features can be defined for which is shown to be quite successful too [4,8].

The use of *syntactic features* is governed by the idea that authors tend to unconsciously use similar syntactic patterns. Information related to the structure of the language is obtained by an in-depth syntactic analysis of the text, usually using some sort of an NLP tool. A single text is characterized by the presence

and frequency of certain syntactic structures. Syntax-based features were introduced in [19], where the rewrite rules frequencies were utilized. Stamatatos et al. [17] used noun, verb and prepositional phrase frequencies. Using a Part-of-speech (POS) tagger one can obtain POS tags and POS tag n-gram frequencies. Using such features excellent results can be achieved [6,10,11,12]. Koppel et al. [10] show that the use of grammatical errors and informal styling (e.g., writing sentences in capital letters) as text features can be useful in authorship attribution.

Our work is based on the composition and evaluation of various afore-mentioned text representation features. We use different character, lexical and syntactic features and adapt them for the use with the Croatian language. We use punctuation marks and vowels frequency as character features. Word length, sentence length and function words frequencies are used as lexical features. For the syntax-based features we use those relatively similar to POS tag and POS tag n-grams frequencies.

3 Text Representation

When constructing an authorship attribution system, the central issue is the selection of sufficiently discriminative features. A feature is discriminative if it is common for one author and rare for all the others. Due to the large number of authors some complex features are very useful if their distribution is specific to each author. Moreover, as the texts from dataset greatly differ in length and topic, it is necessary to use the features independent of such variations. If the features were not independent of such variation that would most certainly reduce generality of system's application and could lead to a decrease of accuracy (e.g., relating author to concrete topic or terms). In the following subsections we will describe different features used.

3.1 Function Words

Function words, such as adverbs, prepositions, conjunctions, or interjections, are words that have little or no semantic content of their own. They usually indicate a grammatical relationship or a generic property [21]. Although one would assume that frequencies of some of the less used function words would be useful indicator of authors style even the frequencies of more common function words can adequately distinguish between the authors. Due to the high frequency of the function words and their significant roles in the grammar, the author usually has no conscious control over their usage in a particular text [1]. They are also topic-independent. Therefore, function words are good indicators of the author's style.

It is difficult to predict whether these words will give equally good results for different languages. Moreover, despite the abundance of research in this field, due to various languages, types and sizes of the texts, it is currently impossible to conclude if these features are generally effective [21].

In addition to function words, in this work we also consider frequencies of auxiliary verbs and pronouns. Their frequencies might be representative of the authors style. This makes the set of totally 652 function words that were used.

3.2 Idf Weighted Function Words

The use of features based on function word frequencies often implies the problem of determining how important, in terms of discrimination, a specific given function word is [6].

To cope with this problem we used a combination of L_p normalization of the length and transformation of the function word occurrence frequency, in particular *idf* (inverse document frequency) measure.

Idf measure is defined as [6]

$$F_{idf}(t_k) = \log \frac{n_d}{n_d(t_k)}, \tag{1}$$

where $n_d(t_k)$ is the number of texts from learning data set that contain word t_k and n_d the total number of the texts in that learning data set. The shown measure gives high values for words that appear in a small number of texts and are thus very discriminatory.

As the *idf* measure uses only the information of the presence of a certain function word in the texts ignoring frequency of that word, a word that appears many times in one single text and once in all the others gets the same value as the one which appears once in all of the texts. Therefore it is necessary to multiply the obtained *idf* measure, of the given function word, with the occurrence frequency of that word in the observed text.

3.3 Part-Of-Speech

Next three features we use are morphosyntactic features. To obtain these features, some sort of NLP tool was required. Croatian language is morphologically complex and it is difficult to develop an accurate and robust POS or MSD (Morphosyntactic Description) tagger. We utilized the method given in [20] that uses inflectional morphological lexicon to obtain POS and MSD of each word from the text. However, as large percentage of word-forms are ambiguous and using the given method we cannot disambiguate homographs – words with same spelling but with different meaning – all possible POS and MSD tags for a given word are considered.

Simplest morphosyntactic features we use are features based on POS frequency, similar to the one used in [11]. The given text is preprocessed and the POS of each word in the text is determined. Features are obtained by counting the number of occurrences of different POS tags and then normalized by the total number of words in the text. The used POS tags are adpositions, conjunctions, interjections, nouns, verbs, adjectives, and pronouns. In addition, category "unknown" is introduced for all the words not found in the lexicon (names, places, etc.).

3.4 Word Morphological Category

More complex morphosyntactic features we use take advantage of morphological category of a word. Each word can be described by the set of morphological categories depending on the word's POS: *case*, *gender*, *number* for nouns, *form*, *gender*, *number*, *person* for verbs and *case*, *degree*, *gender*, *number* for adjectives. Moreover, each morphological category can take one of a number of different values, like for example noun can be in nominative, genitive, dative, accusative, vocative, locative, or instrumental case. Features we use are obtained by counting the number of occurrences of different values for each morphological category and then normalizing them by the total number of words in the text. If, for example, a sentence consists of two nouns and an adjective in nominative case then the number of occurrences of nominative case would be equal three and the normalized nominative case frequency would equal one.

3.5 Part-Of-Speech *n*-grams

POS n-grams frequency-based features are features that utilize the idea of word n-grams applied to POS of words in the text. All the words in a given text are replaced by their POS to make a new text representation that is then used to count the number of occurrences of different POS n-grams. The number of occurrences of every single POS n-gram is normalized by the total number of n-grams to make this feature independent of the text length. The POS we use are those given in Subsection 3.3. Since POS n-gram features can produce very large dimensionality, only 3-grams are considered. Example of an POS n-gram on a word 3-gram "Adam i Eva" ("Adam and Eve") is the "noun conjunction noun" trigram.

Further, we investigate the use of n-grams on POS and function words in parallel. The nouns, verbs and adjectives in a given text are replaced by their POS. All the other words, which are considered to be the function words, are left as they were. Therefore the 3-gram "Adam i Eva" transforms to "noun i noun" 3-gram. Frequencies of such n-grams are then used as features. The use of such features is motivated by the idea of capturing the contextual information of function words. Due to many different pronouns, conjunctions, interjections and prepositions that make many different n-grams, frequency filtering is applied – only the frequency of 500 most frequent 3-grams in the training data set is considered. Used dimension reduction method is not optimal, therefore in the future work other methods should be evaluated, such as information gain, χ^2 test, mutual information, maximum relevance, minimum redundancy, etc.

3.6 Other Features

Other features we use are simple character and lexical features: punctuation marks, vowels, words length and sentence length frequencies.

A set of following punctuation marks is used: ".", ",", "!", "?", "'", """, "-", ":", ";", "+", "*". Their appearance in the text is counted and the result is normalized by the total number of characters in the text.

Features based on the frequency of vowel occurrence (a, e, i, o, u) are obtained in an equal manner.

The frequencies of words lengths are obtained by counting the lengths of all the words from a text and then normalizing them by the total number of words in the given text. To enforce consistency, the words longer than 10 characters are counted as if they were 10 characters long.

The sentence length frequency is obtained in a similar procedure. For the same reason as with the words length, sentences longer than 20 words are counted as if they were 20 words long.

All features suggested in this subsection have weak discriminatory power on their own. However, they proved very useful in the combination with other features, as shown in Table 1.

4 Classification

All the features mentioned in this work use some sort of frequency information that makes it possible to represent them by real valued feature vectors. Having features in a form of vectors, for the classification, we used an SVM (Support Vector Machine) with radial basis function as the kernel. It is shown that, with the use of parameter selection, linear SVM is a special case of an SVM with RBF kernel [9], which removes the need to consider a linear SVM as a potential classifier. RBF kernel is defined as:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2). \tag{2}$$

Before commencing the learning process and classification with the SVM, we scale the data, real valued feature vectors, to ensure equal contribution of every attribute to the classification. Components of every feature vector are scaled to an interval $[0, 1]$.

For the application of SVM classifier in practical problems it is reasonable to consider SVM with soft margins defined by parameter C, as in [3]. Parameter C together with γ used in RBF kernel completely define an SVM model. The search for the appropriate parameters (C, γ), i.e., model selection, is done by the means of cross-validation: using the 5-fold cross-validation on the learning set parameters (C, γ) yielding the highest accuracy are selected and the SVM classifier is learned using them. The accuracy of classification is measured by the expression:

$$acc = \frac{n_c}{N}, \tag{3}$$

where n_c is the number of correctly classified texts and N is the total number of texts. Parameters (C, γ) that were considered are: $C \in \{2^{-5}, 2^{-4}, \cdots, 2^{15}\}$, $\gamma \in \{2^{-15}, 2^{-14}, \cdots, 2^{3}\}$ [2].

5 Data Set

We used three different data sets in our experimental evaluation. First we experimented using an on-line archive of *proofread* articles (journals) from a daily Croa-

tian newspaper "Jutarnji list," available at `http://www.jutarnji.hr/komentari/`. The data set consists of 4571 texts written by 25 different authors. The texts are not evenly distributed among authors – numbers vary from 14 to 1285 texts per author with an average of 182 texts per author. The articles also differ in size – the lowest average number of words in the text per author is 315 words, and the highest average is 1347 words. An average number of words in a text per author is 717 words. Considering this analysis, we can conclude that the used data set is very heterogeneous. Since the writing topics in these articles tend to be time-specific, to avoid the overfitting, we split the set by dates – 20% of the newest articles of each author are taken for testing (hold-out method).

The second data set we used is a collection of on-line blogs. Blog authors were carefully selected: only those authors with more than 40 posts and posts longer than 500 words were considered. In addition, special attention was paid to select only posts that contain original information as authors often tend to quote other sources like news, books, other authors, etc. As a result dataset consisting of 22 authors with a total of 3662 posts was obtained. An average number of words in a post is 846 words. For the testing purposes we split the data set by dates using the same technique that was applied to the articles.

The third data set is comprised of classics of Croatian literature. We have selected 52 novels and novellas from 20 different authors and divided them by chapters, treating each chapter as a separate text (the same approach was used by Argamon et al. [1]). All chapters longer than 4000 words were further divided. That resulted with a total of 1149 texts, with an average of 2236 words in a text. The lowest average number of words in a text per author is 1407 words and the highest average is 3652 words. We have randomly selected 20% of the texts from each author for the test set, leaving the rest as a training set.

The last data set used is a set of Internet forum (message board) posts from `http://www.forum.hr`. We have selected all the posts from the threads in *Politics* category. Initially there were 2269 different authors with a total of 64765 posts. Subsequently we have chosen only those authors that on average have more than 100 words per post and more than 64 posts in total. As a result, the data set consisting of 19 authors and 4026 posts was obtained. Twenty percent of the newest posts of each author were taken as the test set.

6 Evaluation

Classification success is measured by the ratio of correctly classified texts and the total number of texts in the training set i.e., micro average accuracy. First we tested all the features separately and then the same features in various combinations. Results of the evaluation on articles, blogs, and books are shown in Table 1.

Total accuracy does not explain the behavior of the classifier for every class by itself. Therefore, precision and recall are calculated for each class and then used to calculate the total macro F_1 measure.

Table 1. Evaluation of different features using accuracy and F_1 measures [%]

Features	Number of Features	Newspapers		Blogs		Books	
		Acc	F_1	Acc	F_1	Acc	F_1
Function Words ($\mathcal{F}$)	652	88.39	87.38	87.21	87.08	97.24	97.12
Idf Weighted Func. Words ($\mathcal{I}$)	652	87.96	86.84	87.10	86.94	97.24	97.11
Word POS ($\mathcal{C}$)	8	44.50	38.18	34.95	29.85	38.97	33.63
Punctuation Marks ($\mathcal{P}$)	11	57.50	52.93	60.64	59.35	76.55	76.31
Vowels ($\mathcal{V}$)	5	30.54	16.24	25.58	21.84	42.41	39.72
Words Length ($\mathcal{L}$)	11	43.19	33.22	38.59	35.11	47.24	44.15
Sentence Length ($\mathcal{S}$)	20	40.49	33.70	28.00	21.79	25.86	20.12
POS n-grams, 1st meth. ($\mathcal{N}_1$)	512	71.29	68.68	56.67	54.05	80.34	79.83
POS n-grams, 2nd meth. ($\mathcal{N}_2$)	500	76.09	72.52	67.03	66.16	91.72	91.45
Word Morph. Category ($\mathcal{M}$)	22	61.17	58.91	58.32	57.04	67.24	66.13
$\mathcal{C}$, $\mathcal{M}$	30	63.17	62.06	59.65	57.98	71.72	71.84
$\mathcal{P}$, $\mathcal{F}$	663	92.41	91.93	88.75	88.69	**98.62**	**98.57**
$\mathcal{F}$, $\mathcal{M}$	674	91.18	90.68	88.53	88.41	97.93	97.89
$\mathcal{F}$, $\mathcal{N}_1$	1164	89.44	88.52	84.45	84.04	97.59	97.54
$\mathcal{F}$, $\mathcal{N}_2$	1152	90.92	90.43	86.55	86.25	97.93	97.80
$\mathcal{F}$, $\mathcal{C}$	660	90.05	89.52	88.20	87.99	97.59	97.49
$\mathcal{I}$, $\mathcal{M}$	674	90.84	90.35	88.53	88.30	97.93	97.88
$\mathcal{N}_1$, $\mathcal{M}$	534	71.38	70.15	62.29	60.20	84.48	83.80
$\mathcal{I}$, $\mathcal{M}$, $\mathcal{C}$	682	91.36	90.89	88.31	88.16	97.93	97.81
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$	674	**93.37**	92.96	90.30	90.22	**98.62**	**98.57**
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{M}$	696	**93.46**	**93.09**	90.85	90.65	98.28	98.23
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{C}$	682	93.28	92.97	89.97	89.87	98.62	98.57
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{N}_2$	1174	90.31	89.24	87.32	87.21	98.28	98.14
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{M}$, $\mathcal{C}$	704	**93.46**	**93.15**	90.85	90.73	**98.62**	**98.58**
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{M}$, $\mathcal{N}_2$	1196	91.54	90.52	88.09	87.95	98.28	98.14
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{M}$, $\mathcal{C}$, $\mathcal{N}_2$	1204	91.80	90.88	88.64	88.49	97.93	97.81
$\mathcal{P}$, $\mathcal{F}$, $\mathcal{V}$, $\mathcal{L}$	679	93.37	93.04	90.19	90.11	98.62	98.57
$\mathcal{P}$, $\mathcal{C}$, $\mathcal{N}_2$, $\mathcal{M}$	541	85.42	84.11	73.87	73.06	94.14	94.02
$\mathcal{F}$, $\mathcal{M}$, $\mathcal{C}$, $\mathcal{N}_1$	1194	89.62	88.70	86.11	85.69	98.62	98.57
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{L}$	694	92.67	92.18	**91.29**	**91.23**	**99.66**	**99.65**
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{N}_2$, $\mathcal{L}$	542	83.33	81.68	73.98	73.31	95.17	95.06
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{V}$, $\mathcal{L}$	699	93.19	92.85	91.18	91.10	98.62	98.58
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{V}$, $\mathcal{L}$, $\mathcal{M}$	721	93.19	92.88	**91.51**	**91.38**	**98.62**	**98.57**
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{V}$, $\mathcal{L}$, $\mathcal{M}$, $\mathcal{N}_1$	1233	92.41	91.96	89.08	88.71	98.28	98.23
$\mathcal{S}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{V}$, $\mathcal{L}$, $\mathcal{M}$, $\mathcal{N}_1$, $\mathcal{C}$	1241	92.41	91.96	89.75	89.44	98.28	98.23

The SVM parameter selection is very time consuming. Thus, not all of the feature combinations were tested. We focused on the evaluation of the feature combinations that are based on syntactic analysis of the Croatian language and those that are based on function word frequencies as they have proven to be most successful. Furthermore, some of the features like function words and idf weighted function words are very similar and the evaluation of their combination would show no considerable improvements. Also, certain feature combinations made no significant contributions so we did not conduct further evaluation.

As we can see from Table 1, highest accuracies are achieved by using combination of simple features such as function words, punctuation marks, words length, and sentence length frequencies ($\mathcal{F}$, $\mathcal{P}$, $\mathcal{L}$, $\mathcal{S}$). On the other hand, the combinations of syntax-based features are slightly less accurate. If we use the combination of function word and syntax-based features accuracy remains nearly the same as without the use of syntax-based features. We can also see that features based on function words frequencies achieve excellent results on different data sets, regardless of the form of the text, as long as the data set used is com-

prised of texts that are sufficiently large. By increasing the average text length we can see that classification results get higher. To further evaluate the impact of the text length on performance of features used, we have performed experiments on much shorter texts. We used forum posts data set, but the results were not satisfying. The highest accuracy of 27% was achieved by using punctuation marks frequencies.

7 Conclusion

It is shown that the authorship attribution problem, when applied to morphologically complex languages, such as the Croatian language, can be successfully solved using a combination of some relatively simple features. Our results with 91%, 93% and 99% accuracy are quite notable considering the fact that different heterogeneous data sets were used. Evaluation on different data sets shows that the same feature combination based on function words, punctuation marks, word length, and sentence length frequencies, achieve the highest results. Therefore, we conclude that those features are the most suitable for use in the task of authorship attribution. In the similar work of Uzuner et al. [18] on a smaller set of authors, fairly larger texts and with the use of different classifier nearly the same results and conclusions are obtained. The direct comparison of the results has proved very difficult due to the different types of data sets used and different number of authors considered. However, our results fall within the interval of previously reported results, which range from 70% to 99% [1,4,8,12,17,18].

In addition, there are no other reported methods or results for the Croatian language, nor any of the related South Slavic languages, so our work presents a basis for further research.

In future work the problems with homography should be resolved in order to get more accurate results of syntax-based features. Features based on word and character n-grams, suggested in [4,8,14], should be compared to features used in this work. Also, comparison to semantic based features should be conducted. It would be interesting to investigate other features that are able to cope with short texts such as Internet forum posts, e-mails, on-line conversations, or text messages.

Acknowledgement

This research has been supported by the Croatian Ministry of Science, Education and Sports under the grant No. 036-1300646-1986. The authors would like to thank Bojana Dalbelo Bašić and Jan Šnajder for their helpful suggestions.

References

1. Argamon, S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. Proceedings of ACH/ALLC 2005 (2005)

2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3) (1995) 273–297
4. Coyotl-Morales, R., Villaseñor-Pineda, L., Montes-y Gómez, M., Rosso, P., del Lenguaje, L.: Authorship attribution using word sequences. Lecture Notes in Computer Science **4225** (2006) 844
5. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM Sigmod Record **30**(4) (2001) 55–64
6. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Applied Intelligence **19**(1) (2003) 109–123
7. Holmes, D.: Authorship attribution. Computers and the Humanities **28**(2) (1994) 87–106
8. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING. Volume 3. (2003) 255–264
9. Keerthi, S., Lin, C.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation **15**(7) (2003) 1667–1689
10. Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: Proceedings of IJCAI. Volume 3. (2003) 69–72
11. Kukushkina, O., Polikarpov, A., Khmelev, D.: Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission **37**(2) (2001) 172–184
12. Luyckx, K., Daelemans, W.: Shallow text analysis and machine learning for authorship attribution. In: Proceedings of the fifteenth meeting of Computational Linguistics in the Netherlands (CLIN 2004). (2005) 149–160
13. Mendenhall, T.: The characteristic curves of composition. Science (214S) (1887) 237
14. Peng, F., Schuurmans, D., Wang, S., Keselj, V.: Language independent authorship attribution using character level language models. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, Association for Computational Linguistics (2003) 274
15. Stamatatos, E.: Ensemble-based author identification using character n-grams. In: Proceedings of the 3rd International Workshop on Text-based Information Retrieval. (2006) 41–46
16. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology **60**(3) (2009) 538–556
17. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. Computers and the Humanities **35**(2) (2001) 193–214
18. Uzuner, O., Katz, B.: A comparative study of language models for book and author recognition. Lecture Notes in Computer Science **3651** (2005) 969
19. van Halteren, H., Tweedie, F., Baayen, H.: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Computers and the Humanities **28**(2) (1996) 87–106
20. Šnajder, J., Dalbelo Bašić, B., Tadić, M.: Automatic acquisition of inflectional lexica for morphological normalisation. Information Processing and Management **44**(5) (2008) 1720–1731
21. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. Lecture Notes in Computer Science **3689** (2005) 174

List of changes during the revision

- Abstract – modified to include information on different data sets.
- Section 3 Data Set – added 2nd, 3rd, and 4th paragraph.
- Section 6 Evaluation – modified to include information on different data sets, table 1 modified to include new results.
- Section 7 Conclusion – modified 1st and 3rd paragraph.