Automatic Adaptation of Author's Stylometric Features to Document Types

Jan Rygl

Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czech Republic
xrygl@fi.muni.cz

Abstract. Many Internet users face the problem of anonymous documents and texts with a counterfeit authorship. The number of questionable documents exceeds the capacity of human experts, therefore a universal automated authorship identification system supporting all types of documents is needed. In this paper, five predominant document types are analysed in the context of the authorship verification: *books, blogs, discussions, comments and tweets*. A method of an automatic selection of authors' stylometric features using a double-layer machine learning is proposed and evaluated. Experiments are conducted on ten disjunct train and test sets and a method of an efficient training of large number of machine learning models is introduced (163,700 models were trained).

Keywords: authorship verification, feature selection, machine learning, stylome, stylometric features

1 Introduction

The Internet has become an integral part of our lives, it is used in the interpersonal communication (e-mails, forums, chat rooms, etc.) and the promotion of goods and events (product reviews, e-shops, corporate portals,…). Since the Internet is built as an anonymous source of information, many users have faced the problem of anonymous documents and texts with a counterfeit authorship. The sources of serious problems come in different forms, from anonymous threats and menaces [2] through fictitious product reviews and forged e-mail headers to anonymously or spuriously published illegal documents.

As the number of unsigned or unofficially signed documents increases rapidly, it is not possible to solve the problem of the authorship recognition with the capacity of human experts. Instead, automated methods for authorship detection are developed and applied.

Because long texts tend to be consistent and reveal sufficient information about the author's style, most methods have been developed for an analysis of books [11]. The emergence of the field of the authorship identification dates back to the antiquity and was motivated by the literature for a long period of time [10].

Nowadays, the authorship recognition methods are predominantly solved by **machine learning** (ML) methods [14] and stylometric analysis [13] that provide the best results when the input is formed with a sufficiently long and coherent text. However,

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 53–61, 2014.

the current situation of electronic means of communication calls for a universal automated authorship identification system that supports all types of documents. The results of experiments we conducted suggest that the performance of individual stylometric features depends on the length and type of documents. Therefore, we present an algorithm to supplement authorship recognition systems with automatic stylometric feature selection, depending on the type and the content of examined documents.

The fundamental problem of the authorship recognition field is the task of *authorship verification*. All other tasks, such as authorship attribution or clustering, can be converted to the verification problem. The verification problem can be defined in two different forms:

1. *Strict authorship verification:* Confirm or deny a document authorship by a single known author [18, p. 1].
2. *Authorship verification with corpora:* Given a set of documents written by a suspect along with a document dataset collected from the sample population, we want to determine whether or not an anonymous document is written by the suspect [7].

2 Authorship Verification Approaches

For common stylometric approaches to the authorship verification, we are given two texts A, B and two lists of stylometric features $s(A)$ and $s(B)$. Two predominant machine learning approaches to the authorship verification are presented in the following paragraphs:

1. **One classifier per each known author:** We are given $n \gg 1$ texts written by the author of the text A (documents $A_1, A_2, \ldots, A_n$). For each text, the stylomes $s(A_1), s(A_2), \ldots, s(A_n)$ are extracted, and a model M describing the style of the author of texts A_i is trained by ML methods. If the model is trained only from instances representing the same authorship (pairs A_i, A_j), one-class ML approach is used. However, one-class ML performs generally worse than two-class ML [8]. Training instances can be extended by data representing a different authorship (the second class) by adding texts of other authors, distinct from the author of the document A. In both cases, the author's stylomes are used as attributes (value of each stylometric feature is an attribute) for the ML process. If we are given only one text from the known author of A (or B), this method cannot be used.
2. **A classifier from the corpus:** We are given two texts (A and B) and a collection of texts with similar length to the documents A and B written by authors with known authorships: $C^1_{a1}, C^2_{a1}, \ldots, C^n_{a1}, C^1_{a2}, \ldots$. Each document can be compared with each other, therefore the problem is defined as a two-class ML (same authorship: $C^m_{a_i}, C^n_{a_i}$; different authorship: $C^m_{a_i}, C^n_{a_j}, a_i \neq a_j$). Training instances are generated from pairs of documents $(C^m_{a_i}, C^n_{a_j})$ in the following manner: stylomes $s(C_{a_i}), s(C_{a_j})$ are extracted and their normalized absolute difference $sim = 1 - |s(C_{a_i}) - s(C_{a_j})|$ represents a similarity vector corresponding to similarity between two texts ($sim \to 0 \equiv$ *same style*, $sim \to 1 \equiv$ *different style*).
   The ML model M is created from the training instances and the similarity of texts A, B is defined as a probability of the same authorship predicted by M:

$sim \equiv M(1 - |s(A) - s(B)|)$. If we are given more texts by the author(s) of texts A and B, the training instances can be extended by pairs A_i–A_j, B_i–B_j (same authorships) and pairs A_i–$C^m_{a_j}$, B_i–$C^m_{a_j}$ (different authorships).

Since specialized webs as discussion forums, blog servers and libraries contain large text collections with known authorships, the second approach was selected for our experiments. The selected ML implementation supports probability estimation $\langle 0, 0.5)$ for different authorships, $\langle 0.5, 1\rangle$ for same authorships.

3 Author's Stylometric Features – Stylome

Each author has a unique vocabulary, popular phrases, stylistic preferences and a bias that characterizes the author. The author's features are important for the authorship recognition if they allow to distinguish the author from the majority of the population, or if the features can be consistently observed in author's texts.

For the purpose of authorship verification, a vector of decimal numbers (*feature vector*) is extracted from each author's document by a *quantification* of these features. Such vector, when averaged through all author's documents, summarizes all stylometric features of one author. The resulting vector is called a *stylome* [1] – it provides information that defines the style of the author.

The stylome can be analysed via statistical and machine learning methods to discover the author's key features (unique and consistent). The values of key features are used as attributes for ML to recognize whether two documents were written by the same author or not.

The selection of features for the stylome is certainly not fixed. In our experiments conducted on Czech texts, the stylome quantification contains the following 14 categories of features (each category contains tens to thousands of features):

1. **Punctuation analysis** – relative frequencies of punctuation marks are counted [4,12]. We have extended the punctuation analysis with the information about relative positions of punctuation marks in sentences (beginning, middle, or end).
2. **Sentence length distribution** *has been applied with some success and justification when used in conjunction with other attribution tests* [3, p. 17]. Again, we use extended version of the method in experiments – instead of comparing raw differences of frequencies, the data are preprocessed not to penalize differences in close sentence lengths.
3. **Syntactic analysis** – the availability of fast and accurate natural language parsers allows for serious research into syntactic stylometry [5]. SET [9] produces 3 types of sentence parsing trees: a dependency format, a constituent format and a hybrid format. The information from the syntactic trees such as the tree depth, the node count and relative frequency of particular non-terminal nodes are used [16].
4. ***N*-gram syntactic analysis** – instead of relative frequency of non-terminal nodes used in the previous method, the relative frequency of the most common syntactic n-grams is used (syntactic n-grams consist of interdependent non-terminal nodes).
5. **Analysis of morphological categories** – morphological tagging and disambiguation are required to perform analysis of morphological categories in the input text.

The Czech tagger *Desamb* was used to annotate texts and extract relative frequencies of morphological categories that are used as features.

6. **Analysis of morphological tags** uses same tools as the analysis of morphological categories. For each word in a corpus, a normalized morphological-tag information is extracted (set of all important tags assigned to the word). The most frequent morphological tags are used as features and their relative frequencies are counted for each text.
7. **Frequency of word classes** uses fourteen most significant features from the analysis of morphological categories.
8. **Frequency of word-class bigrams** – the number of n-grams of all morphological categories is too big for $n \geq 2$, therefore bigrams of word classes are used as features (14^2 combinations).
9. **Frequency of stop words** – the main advantage of stop word analysis is the topic-independence. The differences of relative stop word frequencies in a large corpus and relative stop word frequencies in the examined text are used as attributes. Lemmatized tokens and stop words are used in our implementation.
10. **Word repetition analysis** – in particular, word lemmata are used here instead of words because Czech is an inflectional language. A separate numeric feature is defined for each word class. If a lemma is repeated in a sentence, value of a feature corresponding to the word class of the lemma is increased.
11. **Simpson's vocabulary richness** – if we manage to collect enough data, we can derive the characteristic of the author on the basis of his or her active vocabulary [6, p. 334]. Simpson in [17] presents a low metric score, which indicates high vocabulary richness: $D = \sum (V_i \cdot \frac{i}{N} \cdot \frac{i-1}{N-1})$ where V_i denotes the number of words with frequency i and N is the number of words in the text.
12. **Analysis of typographic errors** – sum of absolute differences of relative frequencies of selected typographic errors. Each frequency is computed as a ratio of error occurrences in the document to the number of characters in the document.
13. **Frequency of emoticons** – common emoticons were divided to three groups (28 positive, 42 negative, 58 neutral) and relative frequencies of each emoticon and group were used as features.
14. **Analysis of usage of capital letters** – 3 categories of words were distinguished: lowercase words, uppercase words, title words. Relative frequencies of categories and bigrams of two words with the same category were also used as features.

4 Experiments

4.1 Methodology

The evaluation of each category of stylometric features and combination of feature categories is required to discover the best subset of all stylometric features, as two high-quality features can describe similar aspects of the text and learning them in one set achieves worse results than two inferior features that inform about unique trait of the author.

For the above described fourteen feature categories, there are in all combinations $\sum_{i=1}^{14} \binom{14}{i} = 16{,}383$ subsets. Learning a standard machine learning model for thousands of features (categories contains tens to thousands features) is time-consuming (up to hours per model), therefore it is impossible to learn a full model for each subset. We decided to use the double-layer ML approach [15] that creates a model for each feature category (the first layer) and predictions of the first layer models are used as instances for a final model (the second layer):

1. For each category c, a model M_c is created from training instances (14 tasks with tens to thousands features). The models output probabilities of the same authorship of a document pair (a training instance). Train and test data are evaluated separately for each model and each training instance (a document pair) is converted to vector of 14 values (each value corresponds to one feature category).
2. For each subset of feature categories, train and test data are filtered to contain only values that correspond to features in the subset. Since all problems are composed of 14 or less attributes, 16 369 models are trained very fast (tens of seconds per model).

4.2 Data Sources

Texts from five data sources were downloaded and used for experiments (all documents were written by Czech authors), see Table 1. Documents were organized into train and test collections (document pairs were randomly generated). For each data source s, two pairs of collections were built:

1. **unbalanced collections** of productive authors with 5 documents by each of 20 authors – $train_s^{20\times5}$ (2 000 diff. inst., 200 same. inst.) and $test_s^{20\times5}$ (210 diff. inst., 210 same. inst.), where the training instances of the different authorship predominate over instances of the same authorship, and
2. **balanced collections** of 10 documents of each of 10 authors – $train_s^{10\times10}$ (495 diff. inst., 495 same. inst.) and $test_s^{10\times10}$ (473 diff. inst., 472 same. inst.), where the training instances are evenly distributed.

Table 1. Data sources used in experiments

type	doc. #	auth. #	avg. doc. length	source
books	801	79	3 970 words	various
blog articles	26 143	3 352	786 words	`http://blog.idnes.cz/`
posts from forums	3 664	101	358 words	`http://www.filosofie.cz/forum/`
comments under blogs	5 519	3 352	151 words	`http://blog.idnes.cz/`
tweets	1 621	20	33 words	`http://www.klaboseni.cz`

4.3 Results

Each feature category (FC) was evaluated on train data using cross-validation and on test data. The accuracy is an average accuracy of ML models using that FC. The score is a normalized ranking (a model with the best accuracy has rank 1, the second rank 2, etc.) of models using that FC:

$$score_{FC} = \sum_{i=1}^{|models|} \frac{1}{rank(model_i)} \qquad \text{if } FC \in model_i;\ \text{else } 0$$

A higher score corresponds to a better feature category. In the following tables divided per document types, the FC score is shown.

Blogs *Sentence length distribution*, *Frequency of stop words* and *Typographic errors* were the most effective methods and were used by the two best models. Models using *Frequency of emoticons*, *Vocabulary richness*, *Syntactic analysis* and *Word repetition analysis* have not achieved good results. The highest model accuracy for blogs on test data was 72.14 %. For detailed results, see Table 2.

Books *Frequency of word classes* and *Word repetition analysis* were the most effective methods and both the best models have used them. Models using *Vocabulary richness* and *Freq. of word-class bigrams* have not achieved good results. The book authorship verification models expectedly provided the highest model accuracy among all document types at 75 % (see Table 3).

Discussions and User Comments *Frequency of word classes*, *Punctuation analysis* and *Typographic errors* were the most effective methods for these short texts and were again used by the two best models. Models using *Word repetition analysis*, *Morphological categories*, *Usage of capital letters*, *N-gram syntactic analysis* and *Syntactic analysis* have not achieved good results. The best model accuracy on test data for this document type was 72.90 %. For detailed results (see Table 4).

Forums *Morphological categories*, *Typographic errors* and *Frequency of emoticons* were the most effective methods, which both were used by the two best models. Models using *Usage of capital letters* and *Frequency of word classes* have not achieved good results. The best Forums model has reached 73.81 % accuracy (see Table 5).

Tweets For the shortest texts in our experiments, authorship verification works best with *Morphological tags*, *Punctuation analysis*, *Frequency of stop words*, *Word repetition analysis*, *Vocabulary richness* and *Usage of capital letters* as the predominant feature categories used by the best two models. We can see that, as the overall accuracy decreases, there are more feature categories that can help to classify the results. Models using *Morphological categories* have not achieved good results. The overall best accuracy on test data for tweets was 66.43 % (see Table 6).

Table 2. Blog analysis

Feature category	Train$^{10\times10}$	Test$^{10\times10}$		Train$^{20\times5}$	Test$^{20\times5}$		Total	
	Acc.	Acc.	Score	Acc.	Acc.	Score	Score	Rank
Usage of capital letters	53.0%	61.53%	4.59	47.0%	65.52%	8.81	13.40	4
Sentence length distribution	54.0%	61.59%	8.30	53.0%	65.21%	7.78	16.08	2
Frequency of emoticons	50.0%	61.42%	5.02	48.0%	65.14%	4.23	9.25	11
Morphological tags	54.0%	61.53%	4.94	51.0%	65.17%	5.07	10.01	9
Morphological categories	63.0%	61.62%	5.22	56.0%	65.44%	3.59	8.81	13
Punctuation analysis	57.0%	61.71%	8.60	55.0%	65.75%	4.27	12.87	5
Frequency of stop words	61.0%	63.11%	10.27	55.0%	66.96%	10.19	20.45	1
Syntactic analysis	53.0%	61.40%	4.48	48.0%	65.19%	4.69	9.17	12
Typographic errors	48.0%	61.50%	7.77	48.0%	65.32%	8.15	15.91	3
N-gram syntactic analysis	50.0%	61.36%	5.24	45.0%	65.25%	5.69	10.93	8
Vocabulary richness	58.0%	61.64%	6.25	58.0%	65.53%	3.24	9.48	10
Word repetition analysis	59.0%	61.37%	4.08	58.0%	65.20%	1.75	5.83	14
Frequency of word classes	63.0%	61.48%	3.47	60.0%	65.66%	8.86	12.33	6
Freq. of word-class bigrams	60.0%	61.94%	6.45	56.0%	65.58%	5.54	11.99	7
Best Accuracy	70.00%	66.23%		76.00%	72.14%			

Table 3. Book analysis

Feature category	Train$^{10\times10}$	Test$^{10\times10}$		Train$^{20\times5}$	Test$^{20\times5}$		Total	
	Acc.	Acc.	Score	Acc.	Acc.	Score	Score	Rank
Usage of capital letters	72.0%	62.30%	2.27	65.0%	71.35%	8.50	10.77	5
Sentence length distribution	72.0%	62.43%	5.94	65.0%	70.59%	5.13	11.07	4
Frequency of emoticons	59.0%	62.32%	3.90	58.0%	70.65%	7.63	11.53	3
Morphological tags	70.0%	62.29%	5.40	65.0%	70.59%	4.88	10.27	6
Morphological categories	70.0%	62.24%	4.12	65.0%	70.29%	1.74	5.87	11
Punctuation analysis	71.0%	62.20%	2.30	64.0%	70.90%	7.07	9.37	8
Frequency of stop words	70.0%	62.29%	2.43	65.0%	70.43%	1.93	4.35	13
Syntactic analysis	69.0%	62.12%	1.43	63.0%	70.57%	6.72	8.15	10
Typographic errors	49.0%	62.32%	6.03	58.0%	70.59%	4.19	10.22	7
N-gram syntactic analysis	70.0%	62.29%	4.82	65.0%	70.59%	4.45	9.27	9
Vocabulary richness	60.0%	62.22%	2.29	57.0%	70.25%	1.66	3.95	14
Word repetition analysis	72.0%	62.56%	8.98	66.0%	71.20%	7.09	16.07	2
Frequency of word classes	71.0%	62.44%	6.70	66.0%	71.56%	9.77	16.46	1
Freq. of word-class bigrams	70.0%	62.04%	0.97	65.0%	70.56%	4.26	5.23	12
Best Accuracy	75.00%	64.87%		69.00%	75.00%			

Table 4. Discussion analysis

Feature category	Train$^{10\times10}$	Test$^{10\times10}$		Train$^{20\times5}$	Test$^{20\times5}$		Total	
	Acc.	Acc.	Score	Acc.	Acc.	Score	Score	Rank
Usage of capital letters	60.0%	68.53%	2.57	57.0%	65.77%	3.23	5.80	13
Sentence length distribution	55.0%	68.47%	3.97	52.0%	65.81%	8.67	12.65	6
Frequency of emoticons	60.0%	68.99%	9.80	59.0%	65.90%	4.29	14.09	4
Morphological tags	60.0%	68.63%	6.08	56.0%	65.58%	4.87	10.95	8
Morphological categories	64.0%	69.02%	4.64	63.0%	65.55%	3.90	8.54	9
Punctuation analysis	67.0%	69.30%	6.74	67.0%	66.85%	9.97	16.71	2
Frequency of stop words	59.0%	68.62%	4.41	57.0%	65.66%	2.35	6.76	12
Syntactic analysis	62.0%	68.48%	2.42	59.0%	65.26%	1.85	4.27	14
Typographic errors	57.0%	68.72%	7.29	62.0%	66.80%	9.91	17.20	1
N-gram syntactic analysis	55.0%	68.50%	3.34	54.0%	65.44%	4.37	7.71	11
Vocabulary richness	58.0%	68.41%	4.57	57.0%	65.56%	6.64	11.21	7
Word repetition analysis	58.0%	68.59%	3.94	60.0%	65.70%	3.93	7.87	10
Frequency of word classes	64.0%	68.78%	6.09	60.0%	65.50%	6.99	13.08	5
Freq. of word-class bigrams	62.0%	68.87%	8.96	63.0%	65.62%	5.75	14.71	3
Best Accuracy	70.00%	72.90%		74.00%	71.43%			

Table 5. Forum analysis

Feature category	Train$^{10\times10}$	Test$^{10\times10}$		Train$^{20\times5}$	Test$^{20\times5}$		Total	
	Acc.	Acc.	Score	Acc.	Acc.	Score	Score	Rank
Usage of capital letters	55.0%	67.93%	2.12	59.0%	66.72%	4.98	7.10	14
Sentence length distribution	55.0%	68.17%	7.82	62.0%	66.51%	4.07	11.89	7
Frequency of emoticons	49.0%	68.12%	5.85	50.0%	66.77%	7.88	13.72	4
Morphological tags	59.0%	68.15%	5.22	59.0%	66.55%	8.00	13.21	6
Morphological categories	59.0%	68.39%	8.33	65.0%	66.95%	8.67	17.01	2
Punctuation analysis	58.0%	68.09%	3.75	55.0%	67.44%	9.55	13.30	5
Frequency of stop words	57.0%	68.58%	8.35	61.0%	66.38%	3.01	11.36	10
Syntactic analysis	59.0%	68.17%	6.77	62.0%	66.55%	2.98	9.76	12
Typographic errors	52.0%	68.95%	10.02	61.0%	67.99%	10.13	20.16	1
N-gram syntactic analysis	52.0%	67.99%	4.78	51.0%	66.57%	6.64	11.43	9
Vocabulary richness	57.0%	68.10%	6.54	55.0%	66.70%	3.67	10.21	11
Word repetition analysis	57.0%	68.56%	8.25	66.0%	66.52%	6.55	14.81	3
Frequency of word classes	59.0%	68.18%	3.30	63.0%	66.66%	4.45	7.74	13
Freq. of word-class bigrams	60.0%	68.51%	9.24	66.0%	66.39%	2.54	11.78	8
Best Accuracy	72.00%	72.48%		81.00%	73.81%			

Table 6. Twitter analysis

Feature category	Train$^{10\times10}$	Test$^{10\times10}$		Train$^{20\times5}$	Test$^{20\times5}$		Total	
	Acc.	Acc.	Score	Acc.	Acc.	Score	Score	Rank
Usage of capital letters	56.0%	62.93%	10.28	62.0%	61.78%	9.27	19.55	1
Sentence length distribution	53.0%	60.63%	6.37	58.0%	60.99%	6.98	13.35	5
Frequency of emoticons	55.0%	60.82%	6.44	53.0%	60.98%	6.18	12.61	6
Morphological tags	49.0%	60.62%	6.02	48.0%	60.74%	5.20	11.22	10
Morphological categories	55.0%	60.58%	2.84	52.0%	60.72%	4.08	6.92	14
Punctuation analysis	57.0%	61.15%	5.47	60.0%	61.29%	6.84	12.31	7
Frequency of stop words	50.0%	60.78%	7.35	51.0%	61.19%	8.82	16.17	2
Syntactic analysis	55.0%	60.38%	1.69	53.0%	61.04%	6.51	8.20	13
Typographic errors	53.0%	60.84%	5.33	54.0%	60.92%	6.09	11.42	8
N-gram syntactic analysis	54.0%	60.61%	4.81	51.0%	60.91%	6.10	10.91	11
Vocabulary richness	55.0%	60.91%	7.17	56.0%	61.43%	7.53	14.70	4
Word repetition analysis	51.0%	61.07%	8.74	52.0%	60.95%	6.88	15.63	3
Frequency of word classes	53.0%	60.57%	2.44	51.0%	61.03%	6.90	9.34	12
Freq. of word-class bigrams	53.0%	60.58%	5.99	49.0%	60.86%	5.26	11.25	9
Best Accuracy	65.00%	66.02%		70.00%	66.43%			

5 Conclusions and Future Work

The experiments indicate that accuracies of stylometric features achieved in cross-validation tests on train data do not correlate with results obtained on test data. If all stylometric features are used, the authorship recognition performs worse than when using only the optimal feature set. Therefore, we recommend to use the double-layer machine learning technique to select the best performing stylometric features.

We have also provided a methodology for selection of the most effective stylometric features for five predominant document types when not enough training data or training time is available.

In the following research, we plan to conduct experiments on other document types and combinations of different document types, which are still a challenge in the field of authorship verification.

Acknowledgements This work has been partly supported by the Ministry of the Interior of CR within the project VF20102014003.

References

1. Walter Daelemans. Explanation in computational stylometry. In: Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pp. 451–462. Springer Berlin Heidelberg, 2013.
2. James R. Fitzgerald. FBI's communicated threat assessment database: History, design, and implementation. *FBI: Law Enforcement Bulletin*, 76:6–9, 2007.
3. J. W. Grieve. Quantitative authorship attribution: A history and an evaluation of technique. Master's thesis, Simon Fraser University, 2005.
4. O. Hilton. *Scientific examination of questioned documents*. Callaghan, 1956.
5. Charles Hollingsworth. Using dependency-based annotations for authorship identification. In: Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 7499 of *Lecture Notes in Computer Science*, pp. 314–319. Springer Berlin Heidelberg, 2012.
6. D. I. Holmes. The Analysis of Literary Style – A Review. *Journal of the Royal Statistical Society*, 148(4):328–341, 1985.
7. Farkhund Iqbal, Liaquat A. Khan, Benjamin C. M. Fung, and Mourad Debbabi. e-mail authorship verification for forensic investigation. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pp. 1591–1598, New York, NY, USA, 2010. ACM.
8. Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In: *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, p. 62, New York, NY, USA, 2004. ACM.
9. Vojtěch Kovář, Aleš Horák, and Miloš Jakubíček. Syntactic analysis using finite patterns: A new parsing system for czech. In: Zygmunt Vetulani, editor, *LTC*, volume 6562 of *Lecture Notes in Computer Science*, pp. 161–171. Springer, 2009.
10. H. Love. *Attributing Authorship: An Introduction*. Cambridge University Press, 2002.
11. Kim Luyckx and Walter Daelemans. Authorship attribution and verification with many authors and limited data. In: *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pp. 513–520, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
12. G. R. McMenamin and D. Choi. *Forensic Linguistics: Advances in Forensic Stylistics*. Crc Press, 2002.
13. A.Q. Morton and S. Michaelson. The Q-Sum Plot. Technical report, Department of Computer Science, University of Edinburgh, 1990. CSR-3-90.
14. Lisa Pearl and Mark Steyvers. Detecting authorship deception: a supervised machine learning approach using author writeprints. *LLC*, 27(2):183–196, 2012.
15. Jan Rygl and Aleš Horák. Authorship Attribution: Comparison of Single-layer and Double-layer Machine Learning. In: Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 7499 of *Lecture Notes in Computer Science*, pp. 282–289. Springer Berlin Heidelberg, 2012.
16. Jan Rygl, Kristýna Zemková, and Vojtěch Kovář. Authorship Verification based on Syntax Features. In: *Proceedings of Sixth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2012.*, pp. 111–119. Tribun EU, 1st ed. Brno (Czech Republic), 2012.
17. Edward H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.
18. Hans van Halteren. Linguistic profiling for author recognition and verification. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.