Authorship Identification of Romanian Texts with Controversial Paternity

Liviu P. Dinu[1,2], **Marius Popescu**[1], **Anca Dinu**[2,3]

University of Bucharest
[1]Faculty of Mathematics and Computer Science,
[2]Centre for Computational Linguistics
[3]Faculty of Foreign Languages and Literature
Academiei 14, 010014 Bucharest, Romania
ldinu@funinf.cs.unibuc.ro, mpopescu@phobos.cs.unibuc.ro, anca_d_dinu@yahoo.com

Abstract

In this work we propose a new strategy for the authorship identification problem and we test it on an example from Romanian literature: did Radu Albala found the continuation of Mateiu Caragiale's novel *"Sub pecetea tainei"*, or did he write himself the respective continuation? The proposed strategy is based on the similarity of rankings of function words; we compare the obtained results with the results obtained by a learning method (namely Support Vector Machines -SVM- with a string kernel).

1. Introduction

The authorship identification problem is a ancient challenge, and almost in every culture there are a lot of disputed works. The problem of authorship identification is based on the assumption that there are stylistic features that help distinguish the real author from any other possibility. Literary-linguistic research is limited by the human capacity to analyze and combine a small number of text parameters, to help solve the authorship problem. We can surpass limitation problems using computational and discrete methods, which allow us to explore various text parameters and characteristics and their combinations. Using these methods van Halteren et al. (van Halteren et al., 2005) have shown that every writer has a unique fingerprint regarding language use. The set of language use characteristics - stylistic, lexical, syntactic - form the human stylom. Computing and analyzing these language use characteristics from various texts, we can solve text authorship problems.

Marcus (Marcus, 1989) identifies four situation in which text authorship is disputed:

- A text attributed to one author seems non-homogeneous, lacking unity, which raises the suspicion that there may be more than one author. If the text was originally attributed to one author, one must establish which fragments, if any, do not belong to him, and who are their real authors.
- A text is anonymous. If the author of a text is unknown, then based on the location, time frame and cultural context, we can conjecture who the author may be and test this hypothesis.
- If based on certain circumstances, arising from literature history, the paternity is disputed between two possibilities, A and B, we have to decide if A is preferred to B, or the other way around.
- Based on literary history information, a text seems to be the result of the collaboration of two authors, an ulterior analysis should establish, for each of the two authors, their corresponding text fragments.

The text characteristics and parameters used to determine text paternity need not have aesthetic relevance. They must be objective, un-ambiguously identifiable, and quantifiable, such that they can be easily differentiated for different authors.

In this paper we used two strategies to investigate one of the most interesting experiments from Romanian literature. The first strategy is based on Support Vector Machines (SVM) with a string kernel (Section 2.). The second one is a new strategy based on the similarity of rankings of function words.

The novelty of our approach (Section 3.) resides in the way we use information given by the function words frequencies. Given a fixed set of function words (usually the most frequent ones), a ranking of these function words according to their frequencies is built for each text; the obtained ranked lists are subsequently used to compute the distance between two texts. To calculate the distance between two rankings we used Rank distance, a metric introduced in (Dinu, 2003) and which was successfully used in various fields as computational linguistics (in investigating the similarity of Romance languages (Dinu and Dinu, 2005)), bioinformatics (the similarity of DNA strings (Dinu and Sgarro, 2006)), or multi-criteria classification (Dinu and Popescu, to appear). Usage of the ranking of function words in the calculation of the distance instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance more robust acting as a filter, eliminating the noise contained in the values of the frequencies.

In the practical side of this project, we tested the upper strategies to address the following situation from Romanian literature. *Mateiu Caragiale*, one of the most important Romanian novelists, died on 1936, at age of 51. In 1929 he begun to works to the novel *"Sub pecetea tainei"*, but unfortunately he died before finishing this novel. Some decades later, in the 70's, a rumor has agitated the Romanian literary world: it seemed that it was founded the last part of the novel *"Sub pecetea tainei"*. Few human experts agreed that the founded text is in concordance with Mateiu's style, and in the next months almost everybody talked about the huge finding. We have to say that the one who claimed that he has found the last part of the novel was an author (Radu Albala) who's literary style was the closest to Mateiu Caragiale, regarding all the successors of Mateiu. When Albala sees that the claimed last part of novel passed the human experts judgement, he stopped the discussions and said that he is the real author of respective text, text which appears latter with the name *"În deal, pe Militari"* . In fact, this was his challenge: to continue the unfinished novel of Mateiu.

In the following we will show that our methods distinguish between the texts of Albala and the texts of Mateiu, and we also show that Albala was closest to Mateiu's style in the first part of its continuation novel.

2. Classification Experiments

The goal of the classification experiments was to see if the style of Albala can be distinguished from the style of Mateiu in a supervised machine learning scenario. In our experiments we followed the usual setting, treating the problem as a binary classification problem. Each one of the two alleged texts, "Sub pecetea tainei", and "În deal, pe Militari" had to be classified as being written by Mateiu (class -1) or by Albala (class $+1$). For training were used all the others works of the two authors. In order to have a balanced training set, in terms of the number of examples for each author and the length of each example (text), we treated each chapter of the Mateiu's novel "Craii de Curtea-Veche" as a separate text. Thus, 5 negative examples (texts written by Mateiu: the 4 chapters of "Craii de Curtea-Veche" and the novella "Remember") and 5 positive examples (texts written by Albala: all the novellas published by Albala excepting "În deal, pe Militari") resulted. In Table 1 the title of each text, its author and its length (in characters) are listed.

As learning method we used Support Vector Machines (SVM) with a string kernel. String Kernels proved to be effective in authorship attribution (Sanderson and Guenter, 2006; Popescu and Dinu, 2007) and because they treat text as characters string they are language independent.

SVM learning algorithm works by embedding the data into a feature space (a Hilbert space), and searching for linear relations in that space. The embedding is performed implicitly, that is by specifying the inner product between each pair of points rather than by giving their coordinates explicitly. Details about SVM can be found in (Taylor and Cristianini, 2004).

The kernel function offers to the SVM the power to naturally handle input data that are not in the form of numerical vectors, such for example strings. The kernel function captures the intuitive notion of similarity between objects in a specific domain and can be any function defined on the respective domain that is symmetric and positive definite. For strings, a lot of such kernel functions exist with many applications in computational biology and computational linguistics (Taylor and Cristianini, 2004).

One of the most natural ways to measure the similarity of two strings is to count how many substrings of length p the two strings have in common. This give rise to the p-spectrum kernel. Formally, for two strings over an alphabet Σ, $s, t \in \Sigma^*$, the p-spectrum kernel is defined as:

$$k_p(s,t) = \sum_{v \in \Sigma^p} \mathrm{num}_v(s)\mathrm{num}_v(t)$$

where $\mathrm{num}_v(s)$ is the number of occurrences of string v as a substring in s [1] The feature map defined by this kernel associate to each string a vector of dimension $|\Sigma|^p$ containing the histogram of frequencies of all its substrings of length p. Taking into account all substrings of length less than p it will be obtained a kernel that is called the *blended spectrum kernel*:

$$k_1^p(s,t) = \sum_{q=1}^{p} k_q(s,t)$$

As in (Popescu and Dinu, 2007), in our experiments we used the blended spectrum kernel. More precisely we used a normalized version of the kernel to allow a fair comparison of strings of different length:

$$\hat{k}_1^p(s,t) = \frac{k_1^p(s,t)}{\sqrt{k_1^p(s,s)k_1^p(t,t)}}$$

The reason for using this kernel is the fact that, in our opinion, similarity of two strings as it is measured by string kernels reflect the similarity of the two texts as it is given by the short words (2-5 characters) which usually are function words, but also are taken into account other morphemes like suffixes ("ing" for example) which also can be good indicators of author's style.

Because the string kernels work at the character level, we didn't need to split the texts in words or to do any preprocessing. The only editing done to the texts was the replacing of sequences of consecutive space characters (space, tab new line, etc.) with only one space character. This normalization was needed in order to not increase or decrease artificially the similarity between texts because of different spacing.

In all the experiments we used a normalized blended spectrum kernel of 5 characters, $\hat{k}_1^5$. The value of 5 was chosen because it proved to be good in previous attribution tests (Popescu and Dinu, 2007), but also because the most important style indicators in a text are function words which usually are short (2-5 characters).

First we did cross validation in order to establish values for parameters ν for SVM. Also the cross validation had the role of estimating the generalization error of learning methods used, or how reliable these methods are. The relative small number of training examples allowed us to use leave one out cross validation which is considered an almost unbiased estimator of generalization error. Leave one out technique consists of holding each example out, training on all the other examples and testing on the hold out example. For value $\nu = 0.7$ we obtained 0% leave one out error for SVM.

[1] Note that the notion of substring requires contiguity. See (Taylor and Cristianini, 2004) for discussion about the ambiguity between the terms "substring" and "subsequence" across different traditions: biology, computer science.

Author	Type	Title	Length (in characters)
Mateiu Caragiale	Chapters of the novel ”Craii de Curtea-Veche”	Întâmpinarea crailor	32153
		Cele trei hagialâcuri	48975
		Spovedanii	58168
		Asfinţitul crailor	67531
	Novellas	Remember	37219
	Unfinished novel	Sub pecetea tainei	62040
Radu Albala	Novellas	Propyläen Kunstgeschichte	21414
		La Paleologu	88701
		Nişte cireşe	17714
		Sclava iubirii	42633
		Femeia de la miezul nopţii	112091
		În deal, pe Militari	19400

Table 1: Texts used in the experiments

Tested on the two texts in the test set, SVM correctly attributed ”În deal, pe Militari” to Albala and ”Sub pecetea tainei” to Mateiu, but the degrees of confidence of the two predictions were different. ”Sub pecetea tainei” was attributed to Mateiu with a probability of 62.56%, while ”În deal, pe Militari” was attributed to Albala with a probability of 50.56%. This very low confidence indicates that in ”În deal, pe Militari” Albala was very close (concerning the style) to Mateiu.

We repeated the above experiment doing a different preprocessing of the texts. Apart from normalizing spaces (as in the previous experiment), we removed all punctuation marks from the texts. All the other settings remained exactly the same as in the previous experiment (the same training set, the same kernel $\hat{k}_1^5$). Again, the leave one out cross validation error was 0% for the SVM parameter $\nu = 0.7$.

Tested on the two texts this time, ”Sub pecetea tainei” was again correctly attributed to Mateiu with a confidence of 66.87%, but ”În deal, pe Militari” was also attributed to Mateiu with a confidence of 58.94%.

The role of punctuation in authorship identification problem was anticipated by Chaski (Chaski, 1996).

The punctuation (especially *,* and *;*) is the one who betrayed on Albala in his challenge to continue the novel of Mateiu, and we can say that the Albala’s stylom is different on the Mateiu’s stylom either on a alternative breathing of comas.

3. Clustering Experiments

Compared with other machine learning and statistical approaches, clustering was relatively rarely used in stylistic investigations. However, few researchers (Holmes et al., 2001; Labbé and Labbé, 2006; Luyckx et al., 2006) have recently proved that clustering can be a useful tool in computational stylistic studies.

An agglomerative hierarchical clustering algorithm (Duda et al., 2001) arranges a set of objects in a family tree (dendogram) according to their similarity. The goal of the clustering experiments was to see how the dendogram of the works of the two authors look like (if the texts belonging to one author are cluster together) and to see where in this family tree are placed the two texts of interest, ”În deal, pe Militari” and ”Sub pecetea tainei”.

In order to work, an agglomerative hierarchical clustering algorithm needs to measure the similarity between objects that have to be clustered, similarity which in its turn is given by a distance function defined on the set of respective objects.

In our experiments we used a new distance measure (Popescu and Dinu, forthcoming) designed to reflect stylistic similarity between texts. As style markers it use the function word frequencies. Function words are generally considered good indicators of style because their use is very unlikely to be under the conscious control of the author and because of their psychological and cognitive role (Chung and Pennebaker, 2007). Also function words prove to be very effective in many author attribution studies. The novelty of the distance measure resides in the way it use the information given by the function word frequencies. Given a fixed set of function words (usually the most frequent ones), a ranking of these function words according to their frequencies is built for each text; the obtained ranked lists are subsequently used to compute the distance between two texts. To calculate the distance between two rankings we used *Rank distance* (Dinu, 2003), an ordinal distance tightly related to the so-called *Spearman’s footrule* (Diaconis and Graham, 1977).

Usage of the ranking of function words in the calculation of the distance instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance measure more robust acting as a filter, eliminating the *noise* contained in the values of the frequencies. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more relevant than the fact that the respective word appears 349 times in the first text and only 299 times in the second.

Rank distance (Dinu, 2003) is an ordinal metric able to compare different rankings of a set of objects.

A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, \ldots, n$, $\sigma \in S_n$. $\sigma(i)$ will represent the place (rank) of the object i in the ranking. The Rank distance in this case is simply the distance induced by

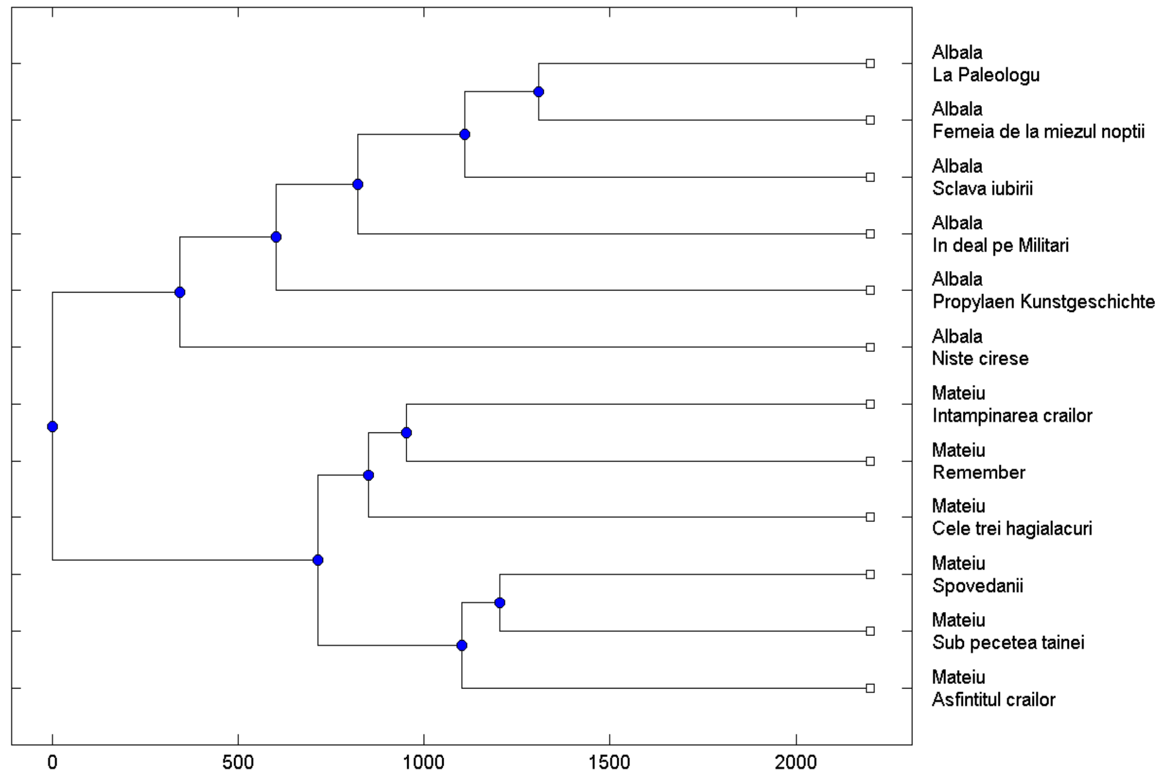

Figure 1: Dendogram of the works of Mateiu and Albala

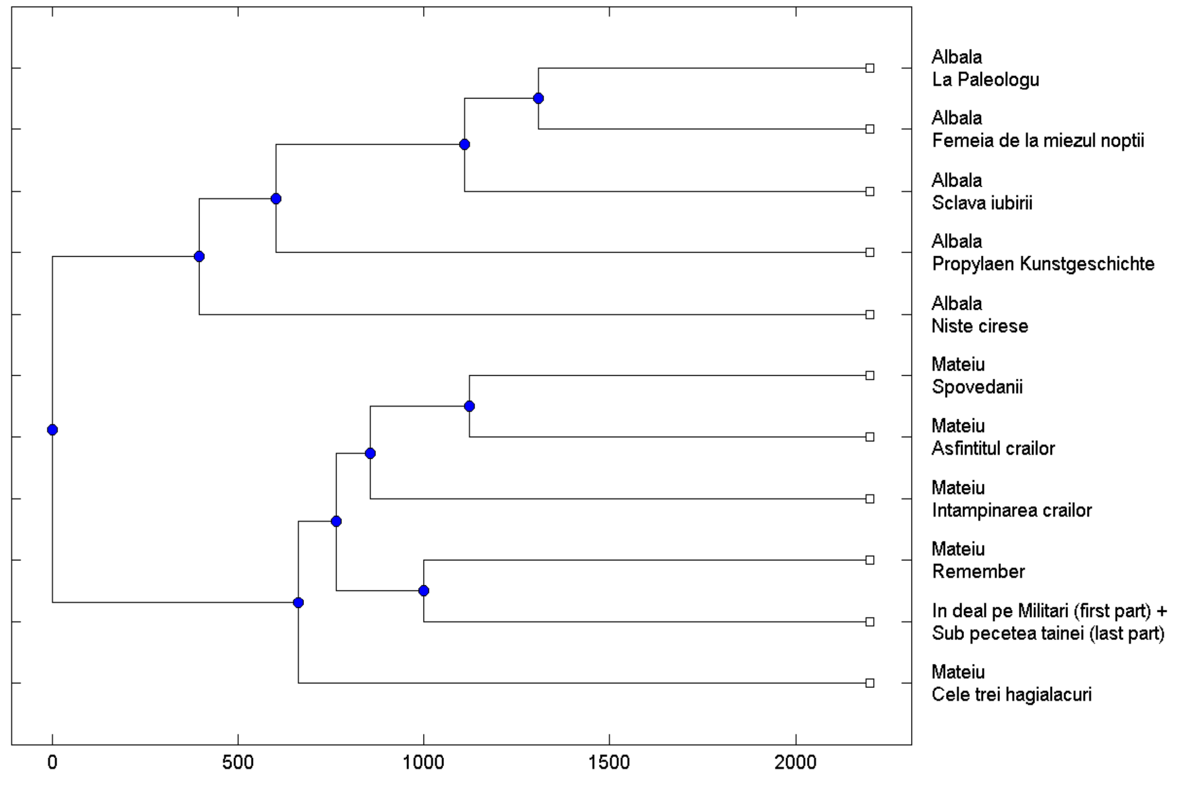

Figure 2: Dendogram of the works of Mateiu and Albala

L_1 norm:

$$D(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| \quad (1)$$

This is a distance between what is called full rankings. However, in real situations, the problem of *tying* arises, when two or more objects claim the same rank (are ranked equally). For example, two or more function words can have the same frequency in a text and any ordering of them would be arbitrary.

The Rank distance allocates to tied objects a number which is the average of the ranks the tied objects share. For instance, if two objects claim the rank 2, then they will share the ranks 2 and 3 and both will receive the rank number $(2 + 3)/2 = 2.5$. In general, if k objects will claim the same rank and the first x ranks are already used by other objects, then they will share the ranks $x + 1, x + 2, \dots, x+k$ and all of them will receive as rank the number: $\frac{(x+1)+(x+2)+\dots+(x+k)}{k} = x + \frac{k+1}{2}$. In this case, a ranking will be no longer a permutation ($\sigma(i)$ can be a non integer value), but the formula (1) will remain a distance (Dinu, 2003).

Rank distance can be used as a stylistic distance between texts in the following way:

First a set of function word must be fixed. The most frequent function words may be selected or other criteria may be used for selection. In all our experiments we used a set of 120 most frequent Romanian function words.

Once the set of function words is established, for each text a ranking of these function word is computed. The ranking is done according to the function word frequencies in the text. Rank 1 will be assigned to the most frequent function word, rank 2 will be assigned to the second most frequent function word, and so on. The ties are resolved as we discussed above. If some function words from the set don't appear in the text, they will share the last places (ranks) of the ranking.

The distance between two texts will be the Rank distance between the two rankings of the function words corresponding to the respective texts.

Having the distance measure, the clustering algorithm initially assigns each object to its own cluster and then repeatedly merges pairs of clusters until the whole tree is formed. At each step the pair of nearest clusters is selected for merging. Various agglomerative hierarchical clustering algorithms differ in the way in which they measure the distance between clusters. Note that although a distance function between objects exists, the distance measure between clusters (set of objects) remains to be defined. In our experiments we used the *complete linkage* distance between clusters, the maximum of the distances between all pairs of objects drawn from the two clusters (one object from the first cluster, the other from the second).

We used the clustering with Rank distance to cluster exactly the same texts that we used in classification experiments (Table 1). The resulted dendrogram is shown in Figure 1. It is easy to see that Mateiu's workss and Albala's works are clustered into two distinct groups, and, that the two investigated texts are placed in their corresponding branch.

To see if indeed Albala wanted to write in the matein style, we made an ad-hoc experiment: we concatenated the last part of the novel *”Sub pecetea tainei”* with the first part of the *”În deal, pe Militari”* and used this artificial text in the same experiment as the previous one. The result was (see Figure 2) that the new ad-hoc text is placed in the Mateiu's branch. Conclusion is that Albala wrote in the beginning of the *”În deal, pe Militari”* as Mateiu, but his concentration decreased towards the end of the novel and eventually his stylom was detectable.

4. Conclusions

The authorship identification problem is a ancient challenge, and almost in every culture there are a lot of disputed papers. In this work we proposed a new strategy for the authorship identification problem and we have tested on an example from Romanian literature: Radu Albala found the continuing of Mateiu Caragiale's novel *”Sub pecetea tainei”*, or he write himself the respective continuing? The answer is that Albala write himself the continuing.

Acknowledgements

Research supported by MEdC-ANCS, PNII-Idei, project 228. We want to thank to Ştefan Agopian for fruitful discussions and to Humanitas Ed. for the electronic source of Albala's texts.

5. References

Carole E. Chaski. 1996. Linguistic methods of determining authorship. In *National Institute of Justice Research Seminar. 49th American Academy of Forensic Sciences Meeting*, Nashville, TN.

Cindy K. Chung and James W. Pennebaker. 2007. The psychological function of function words. In K. Fiedler, editor, *Social communication: Frontiers of social psychology*, pages 343–359. Psychology Press, New York.

P. Diaconis and R.L. Graham. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(2):262–268.

Anca Dinu and Liviu Petrisor Dinu. 2005. On the syllabic similarities of romance languages. In *CICLing, Lecture Notes in Computer Science*, pages 785–788.

Liviu Petrisor Dinu and Marius Popescu. to appear. A multi-criteria decision method based on rank distance. *Fundamenta Informaticae*.

Liviu Petrisor Dinu and Andrea Sgarro. 2006. A low-complexity distance for dna strings. *Fundamenta Informaticae*, 73(3):361–372.

Liviu Petrisor Dinu. 2003. On the classification and aggregation of hierarchies with different constitutive elements. *Fundamenta Informaticae*, 55(1):39–50.

R. O. Duda, P. E. Hart, and D. G. Stork. 2001. *Pattern Classification (2nd ed.)*. Wiley-Interscience Publication.

David I. Holmes, Lesley J. Gordon, and Christine Wilson. 2001. A widow and her soldier: Stylometry and the american civil war. *Literary and Linguistic Computing*, 16(4):403–420.

Cyril Labbé and Dominique Labbé. 2006. A tool for literary studies: Intertextual distance and tree classification. *Literary and Linguistic Computing*, 21(3):311–326.

Kim Luyckx, Walter Daelemans, and Edward Vanhoutte. 2006. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of LREC-2006, the fifth International Language Resources and Evaluation Conference*, pages 30–35.

Solomon Marcus. 1989. *Inventie si descoperire*. Ed. Cartea Romaneasca, Bucuresti.

Marius Popescu and Liviu P. Dinu. 2007. Kernel methods and string kernels for authorship identification: The federalist papers case. In *Proceedings of the International Conference RANLP - 2007 (Recent Advances in Natural Language Processing)*, Borovets, Bulgaria, September.

Marius Popescu and Liviu P. Dinu. forthcoming. Rank distance as a stylistic similarity.

Conrad Sanderson and Simon Guenter. 2006. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Sydney, Australia, July. Association for Computational Linguistics.

John S. Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Hans van Halteren, M. Haverkort, H. Baayen, A. Neijt, and F. Tweedie. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12:65–77.