Overview of the Author Identification Task at PAN 2013

Patrick Juola[1] and Efstathios Stamatatos[2]

[1]Duquesne University
juola@mathcs.duq.edu

[2]University of the Aegean
stamatatos@aegean.gr

Abstract. The author identification task at PAN-2013 focuses on author verification where given a set of documents by a single author and a questioned document, the problem is to determine if the questioned document was written by that particular author or not. In this paper we present the evaluation setup, the performance measures, the new corpus we built for this task covering three languages and the evaluation results of the 18 participant teams that submitted their software. Moreover, we survey the characteristics of the submitted approaches and show that a very effective meta-model can be formed based on the output of the participant methods.

1 Introduction

Authorship attribution is an important problem in many areas including information retrieval and computational linguistics, but also in applied areas such as law and journalism where knowing the author of a document (such as a ransom note) may be crucial to save lives. The most common framework for testing candidate algorithms is a closed-set text classification problem: given known sample documents from a small, finite set of candidate authors, which if any wrote a questioned document of unknown authorship? [16, 33] It has been commented, however, that this may be an unreasonably easy task [22]. A more demanding problem is author verification where given a set of documents by a single author and a questioned document, the problem is to determine if the questioned document was written by that particular author or not [24]. This may more accurately reflect real life in the experiences of professional forensic linguists, who are often called upon to answer this kind of question. Interestingly, every author identification problem with multiple candidate authors can be transformed to a set of author verification problems.

The author identification task at PAN 2013 introduced several new aspects this year. The problem was framed differently this year, using the idea of the "fundamental problem of authorship attribution" as framed by Koppel *et al.* [23], a reframing that supported a new software submission paradigm. The corpus incorporated a substantial multilingual element, including both resource-rich (English, Spanish) and resource-poor (Greek) natural languages. Despite this new framework,

participation remained robust, with 18 participants and 16 notebook submissions, as detailed in the following sections.

2 Relevant Work

Authorship attribution has been a regular task at PAN/CLEF for a number of years: PAN 2011 [1] focused on English language email extracted from the Enron corpus; PAN 2012 [17] focused on a more eclectic set of problems of various types ranging from authorship attribution to document segmentation by author.

For readers unfamiliar with this problem, a brief introduction may be in order. In the all-too-common case where the authorship of a document is important, but unknown, it may be possible to make an educated guess by examining the writing style of the document in question. (As an oversimplified example, if the document uses the British spelling of the word "honour," the writer is likely to be from the UK as opposed to the USA.) This might be important, for example, in the case of a disputed will (where the deceased is from the USA, but the will uses British spellings throughout). This kind of determination can be made "by hand" via skilled linguistic analysis [4, 11] or by computer as in this evaluation. The basis on which such decisions can be made varies widely and is the study of much active research, and the reader is referred to several recent surveys [16, 21, 33].

3 Evaluation Setup

Traditionally, authorship attribution is divided into two types of problems, open- and closed-class problems, with authorship verification being treated as a subtype and special case of the open-class problem. Perhaps obviously, authorship attribution requires a document or documents of unknown authorship (the unknown or questioned documents). In order for analysis to be practical, there must also be documents of known authorship to test against. In the closed-class problem, there are several candidate authors, each represented by a set of training data, and one of these candidate authors is assumed to be the author (i.e, the set of potential authors is a closed set). In the open-class problem, the set of potential authors is an open class, and "none of the above" is a potential answer. Authorship verification is the special case where the set of candidate authors is a singleton, i.e. there is only one candidate, and either he wrote the unknown document(s) or "someone else" did, where "someone else" could be anyone else in the universe.

The question of the appropriate type of problem to use has been controversial. In a modern forensic context, the police have usually done a preliminary investigation before they settle on a set of candidate suspects [11] and thus have narrowed the problem down to an effectively closed set of people with opportunity. In a more literary or historic context, there is usually no way to exclude the possibility of a previously unknown author and so an open set is often more appropriate. Closed-class problems are generally considered to be easier, and partly for this reason, previous evaluations have concentrated on closed-class problems [15, 17].

This year represents a departure from this precedent, as we focus on authorship verification, or what Koppel *et al.* [23] have called the "fundamental problem" in authorship attribution: *Given two documents, are they by the same author?* There is an elegance about this formulation, but it also represents possibly the most difficult formulation of the problem as it contains the minimum extra information upon which an analysis can rely. Discussions of this issue at and after the Authorship Attribution Workshop at Brooklyn Law School (October, 2012) suggested that this framing may be too difficult to solve at present technologies, especially at with realistic amounts of training data. For this reason, we focused on a variant of the fundamental problem: *Given a set of documents (no more than 10, possibly only one) by the same author, is an additional (out-of-set) document also by that author?*

This framework has several advantages, most notably that evaluation is relatively straightforward as each "problem" has a simple yes/no answer and that each problem can be represented relatively simply in a computational framework. This made it easier to incorporate the second major innovation of this iteration of the evaluation, the use of software-only submissions. In contrast to previous years, participants were asked to submit executable programs conforming to a simple command-line interface and output in a specific format that can be automatically evaluated. (Readers familiar with the ACM International Collegiate Programming Contest will be familiar with this paradigm). Submitted programs were run and evaluated in the TIRA[1] platform [10]. Among other advantages, this enables us to "keep the contest open"; if someone has a brilliant idea in 2015, we hope they will be able to use the identical setup to submit and test that idea, hopefully outperforming 2013's winner.

Beyond the binary yes/no answers, it was also possible to leave some problems unanswered. In addition, the participants could optionally produce a confidence score, namely a real number in the set [0,1] inclusive where 1.0 means that it is absolutely sure that the questioned document was written by the examined author and 0 means the opposite.

4 Evaluation Corpus

The corpus we built for the author identification task of PAN-2013 covers three languages: English, Greek, and Spanish. For each language there is a set of problems, where one problem comprises a set of documents of known authorship by the same author and exactly one document of questioned authorship. All the documents within a problem are in the same language and placed in a separate folder. The language information was encoded in the problem label (i.e., folder name) so that it is possible to apply appropriate models per language without the need of language identification techniques.

The training corpus comprised 10 problems in English, 20 problems in Greek and 5 problems in Spanish. On the other hand, the evaluation corpus was balanced over the three languages comprising 30 problems in English, 30 problems in Greek and 25 problems in Spanish. A part of the latter was used in the early-bird evaluation stage,

[1] http://tira.webis.de

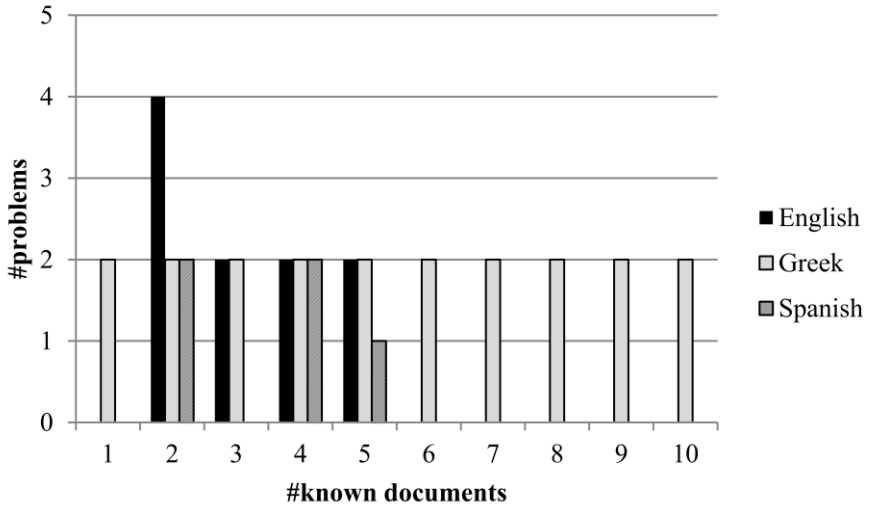

Figure 1. Distribution of known documents over the problems of the training corpus.

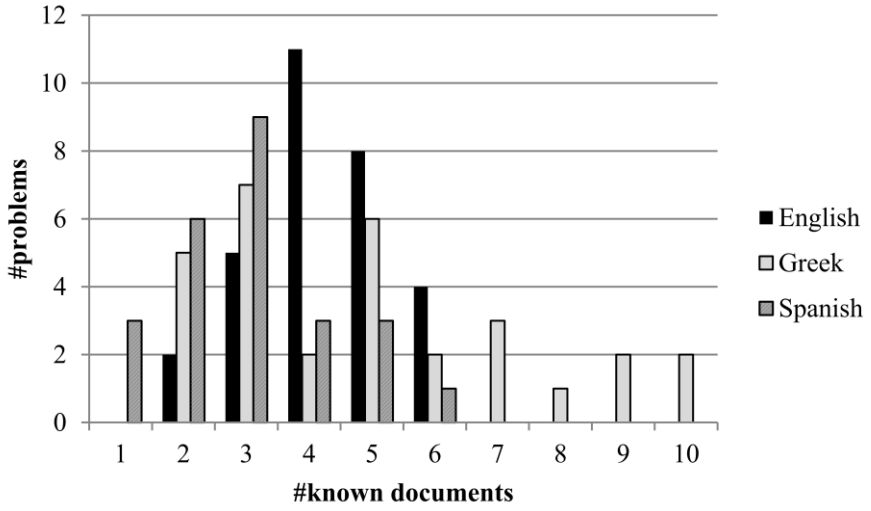

Figure 2. Distribution of known documents over the problems of the evaluation corpus.

that is 20 problems in English, 20 problems in Greek and 15 problems in Spanish. In all cases, the distribution of positive and negative problems in each corpus (and every language-specific sub-corpus) was balanced.

The English part of the corpus (collected by Patrick Brennan of *Juola & Associates*) consists of extracts from published textbooks on computer science and related disciplines, culled from an on-line repository. This particular genre was chosen in part because it represents a relatively controlled universe of discourse and

also a relatively unstudied genre compared with more commonly analyzed genres such as fiction or news reportage. A pool of 16 authors was selected and their works were collected. Each test and training document was around 1,000 words each and collected by hand from the larger works. Formulas and computer code was removed. Beyond the overall genre of "textbooks regarding IT or computer science", some of the paired documents are members of a very narrow genre (e.g. textbooks regarding Java programming) while others are more divergent (e.g. Cyber Crime vs. Digital Systems Design); the intention was to make the task more difficult and to curb a simple reading of the documents for content in order to guess authorship.

The Greek part of the corpus comprises newspaper articles published in the Greek weekly newspaper TO BHMA[2] from 1996 to 2012. Initially, a pool of more than 800 opinion articles by about 100 authors was downloaded. The length of each article is at least 1,000 words. All HTML tags, scripts etc. as well as the title/subtitles of the article and author names were removed semi-automatically. Based on this collection of documents, a set of author verification problems was formed. In each problem, we included texts that had strong thematic similarities indicated by the occurrence of certain keywords. In addition, to make the task more challenging, we applied a stylometric analysis based on a character 3-gram representation and the dissimilarity measure d_1 proposed in [32] to detect stylistically similar or dissimilar documents. Hence, in problems where the true answer is positive (the questioned document was written by the author of the known documents) the unknown document was selected to have relatively high dissimilarity from the other known documents. On the other hand, in problems where the true answer is negative the unknown document (by a certain author) was selected to have relatively low dissimilarity from the known documents (by another author). Therefore, beyond similarities in genre, theme, and date of writing, there also stylistic relationships in within-problem documents of the Greek sub-corpus. This makes the Greek part of the evaluation corpus more challenging especially for verification methods based on CNG and variants [33].

The Spanish part of the corpus (collected in part by Sheila Queralt of Universitat Pompeu Fabra and by Angela Melendez of Duquesne University) consisted of excerpts from newspaper editorials and short fiction.

Figures 1 and 2 show the distribution of known authorship documents per language for the training and evaluation corpora, respectively. In the training corpus, the English and Spanish parts include problems with no more than 5 known documents. On the other hand the Greek part covers the range of 1-10 known documents in a balanced way. In the evaluation corpus, the English and Spanish parts include problems with no more 6 known documents. The majority of the problems comprise 4-5 known documents for English and 2-3 known documents for Spanish. The Greek part again covers the range 1-10 of known documents while the majority of problems include 2-5 known documents.

[2] http://www.tovima.gr

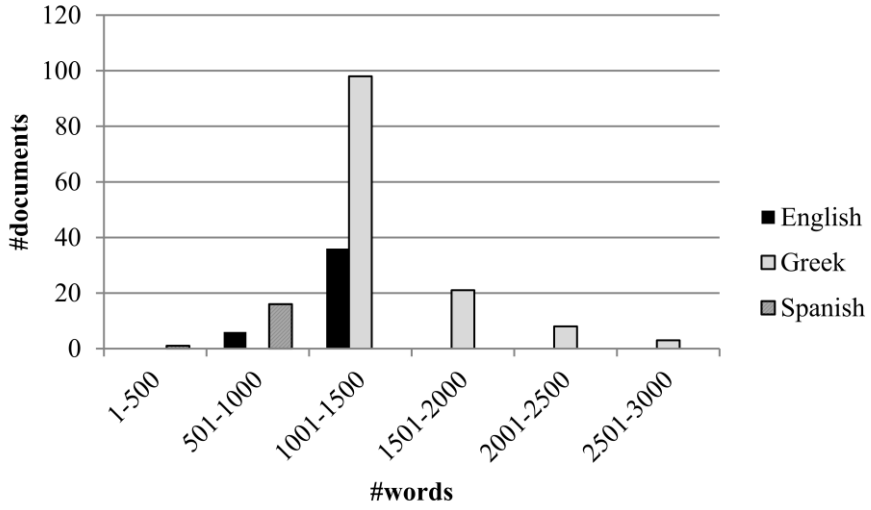

Figure 3. Text-length distribution of the training corpus.

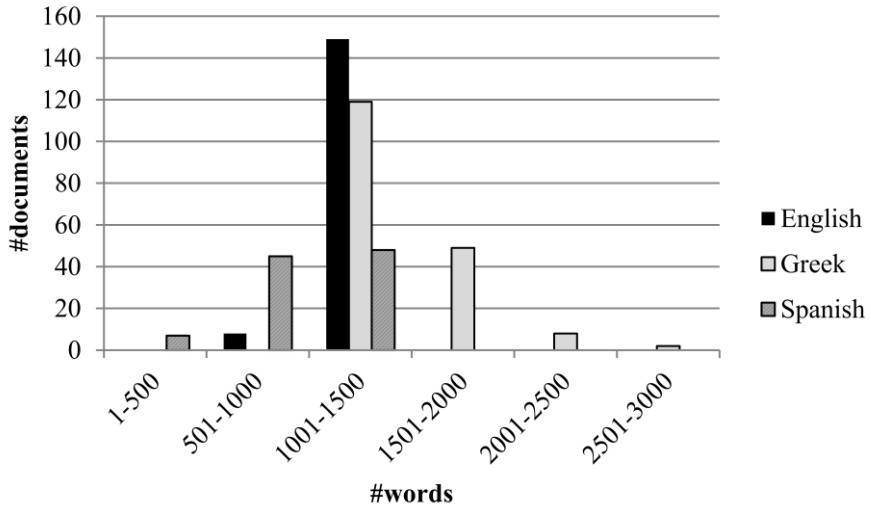

Figure 4. Text-length distribution of the evaluation corpus.

From another perspective, Figures 3 and 4 show the distribution of text-length (in terms of words) over the evaluation documents per language for the training and evaluation corpora, respectively. In both cases, the majority of the documents comprise 1,000-1,500 words while some longer documents are included in the Greek part and some shorter documents in the Spanish part. The distribution of English and Greek parts is similar in training and evaluation corpora. The Spanish part of the evaluation corpus includes longer texts in comparison to the training corpus.

5 Performance Measures

PAN-2013 participants were asked to provide a simple "yes/no" binary answer for each problem of the author identification task. In case their method was not confident enough for some problems, they could leave the problem unanswered. To evaluate the output of their software, we used the following measures:

$$Recall = \#correct\_answers / \#problems$$

$$Precision = \#correct\_answers / \#answers$$

Note that in case a participant answers all the problems, these two measures are equal. The final ranking was computed by combining these measures via F_1 for the whole evaluation corpus comprising all three languages. That way, a method that can only deal with a certain language will be ranked very low.

In addition, to evaluate the participants that also submitted a score (a real number in the set [0,1] inclusive where 1 indicates a confident positive answer and 0 indicates a confident negative answer) we used *Receiver-Operating Characteristic* (ROC) curves and the area under the curve (AUC) as a single measure. ROC curves provide a more detailed picture over the ability of the author verification methods to assign appropriate scores to their answers [6]. For the calculation of ROC curves, any missing answers were assumed to be wrong answers. Again, those participants that can only handle documents of a certain language will produce low AUC scores.

Finally, since we locally run the software submissions, it is possible for first time to have some comparative results between author verification methods with respect to their runtime.

6 Evaluation Results

In total, 18 teams submitted their author verification software. The final evaluation results and the ranking of the participants according to the overall F_1 score as well as their runtime are shown in Table 1. Evaluation results by each one of the three examined languages can be seen in Tables 2, 3, and 4.

Most of the submissions answered all the problems in the evaluation corpus. Hence, the recall and precision measures are equal. Only two participants (Bobicev [2] and Ghaeini [9]) used the “I don’t know” option in some problems. Moreover, two participants (Veenman&Li [35] and Sorin) provided answers only for the English part of the corpus.

The winning submission [31] is a modification of the recently proposed *Impostors* method [20]. It achieved remarkable performance on English and Greek parts of the corpus. On the other hand, its performance on the Spanish part was moderate. Veenman&Li [35] submitted another very effective approach for English only. The submissions of Halvani *et al.*, [12], Layton *et al*. [25], and Petmanson [29] were also noticeable. Beyond their good performance, the former two required very low runtime.

Table 1. Overall results, runtime, and ranking of submissions.

Rank	Submission	F_1	Precision	Recall	Runtime
1	Seidman [31]	0.753	0.753	0.753	65476823
2	Halvani *et al.* [12]	0.718	0.718	0.718	8362
3	Layton *et al.* [25]	0.671	0.671	0.671	9483
3	Petmanson [29]	0.671	0.671	0.671	36214445
5	Jankowska *et al.* [13]	0.659	0.659	0.659	240335
5	Vilariño *et al.* [36]	0.659	0.659	0.659	5577420
7	Bobicev [2]	0.655	0.663	0.647	1713966
8	Feng&Hirst [7]	0.647	0.647	0.647	84413233
9	Ledesma *et al.* [26]	0.612	0.612	0.612	32608
10	Ghaeini [9]	0.606	0.671	0.553	125655
11	van Dam [5]	0.600	0.600	0.600	9461
11	Moreau&Vogel [27]	0.600	0.600	0.600	7798010
13	Jayapal&Goswami [14]	0.576	0.576	0.576	7008
14	Grozea	0.553	0.553	0.553	406755
15	Vartapetiance&Gillam [34]	0.541	0.541	0.541	419495
16	Kern [19]	0.529	0.529	0.529	624366
	BASELINE	0.500	0.500	0.500	
17	Veenman&Li [35]	0.417	0.800	0.282	962598
18	Sorin	0.331	0.633	0.224	3643942

Table 2. Results on the English part of the evaluation corpus.

Submission	F_1	Precision	Recall
Seidman [31]	0.800	0.800	0.800
Veenman&Li [35]	0.800	0.800	0.800
Layton *et al.* [25]	0.767	0.767	0.767
Moreau&Vogel [27]	0.767	0.767	0.767
Jankowska *et al.* [13]	0.733	0.733	0.733
Vilariño *et al.* [36]	0.733	0.733	0.733
Halvani *et al.* [12]	0.700	0.700	0.700
Feng&Hirst [7]	0.700	0.700	0.700
Ghaeini [9]	0.691	0.760	0.633
Petmanson [29]	0.667	0.667	0.667
Bobicev [2]	0.644	0.655	0.633
Sorin	0.633	0.633	0.633
van Dam [5]	0.600	0.600	0.600
Jayapal&Goswami [14]	0.600	0.600	0.600
Kern [19]	0.533	0.533	0.533
BASELINE	0.500	0.500	0.500
Vartapetiance&Gillam [34]	0.500	0.500	0.500
Ledesma *et al.* [26]	0.467	0.467	0.467
Grozea	0.400	0.400	0.400

Table 3. Results on the Greek part of the evaluation corpus.

Submission	F_1	Precision	Recall
Seidman [31]	0.833	0.833	0.833
Bobicev [2]	0.712	0.724	0.700
Vilariño *et al*. [36]	0.667	0.667	0.667
Ledesma *et al*. [26]	0.667	0.667	0.667
Halvani *et al*. [12]	0.633	0.633	0.633
Jayapal&Goswami [14]	0.633	0.633	0.633
Grozea	0.600	0.600	0.600
Jankowska *et al*. [13]	0.600	0.600	0.600
Feng&Hirst [7]	0.567	0.567	0.567
Petmanson [29]	0.567	0.567	0.567
Vartapetiance&Gillam [34]	0.533	0.533	0.533
BASELINE	0.500	0.500	0.500
Kern [19]	0.500	0.500	0.500
Layton *et al*. [25]	0.500	0.500	0.500
van Dam [5]	0.467	0.467	0.467
Ghaeini [9]	0.461	0.545	0.400
Moreau&Vogel [27]	0.433	0.433	0.433
Sorin	-	-	-
Veenman&Li [35]	-	-	-

Table 4. Results on the Spanish part of the evaluation corpus.

Submission	F_1	Precision	Recall
Halvani *et al*. [12]	0.840	0.840	0.840
Petmanson [29]	0.800	0.800	0.800
Layton *et al*. [25]	0.760	0.760	0.760
van Dam [5]	0.760	0.760	0.760
Ledesma *et al*. [26]	0.720	0.720	0.720
Grozea	0.680	0.680	0.680
Feng&Hirst [7]	0.680	0.680	0.680
Ghaeini [9]	0.667	0.696	0.640
Jankowska *et al*. [13]	0.640	0.640	0.640
Bobicev [2]	0.600	0.600	0.600
Moreau&Vogel [27]	0.600	0.600	0.600
Seidman [31]	0.600	0.600	0.600
Vartapetiance&Gillam [34]	0.600	0.600	0.600
Kern [19]	0.560	0.560	0.560
Vilariño *et al*. [36]	0.560	0.560	0.560
BASELINE	0.500	0.500	0.500
Jayapal&Goswami [14]	0.480	0.480	0.480
Sorin	-	-	-
Veenman&Li [35]	-	-	-

Table 5. Evaluation of real scores (AUC) for the whole corpus and per language.

Rank	Submission	Overall	English	Greek	Spanish
1	Jankowska, *et al*. [13]	**0.777**	**0.842**	0.711	0.804
2	Seidman [31]	0.735	0.792	**0.824**	0.583
3	Ghaeini [9]	0.729	0.837	0.527	**0.926**
4	Feng&Hirst [7]	0.697	0.750	0.580	0.772
5	Petmanson [29]	0.651	0.672	0.513	0.788
6	Bobicev [2]	0.642	0.585	0.667	0.654
7	Grozea	0.552	0.342	0.642	0.689
	BASELINE	0.500	0.500	0.500	0.500
8	Kern [19]	0.426	0.384	0.502	0.372
9	Layton *et al*. [25]	0.388	0.277	0.456	0.429

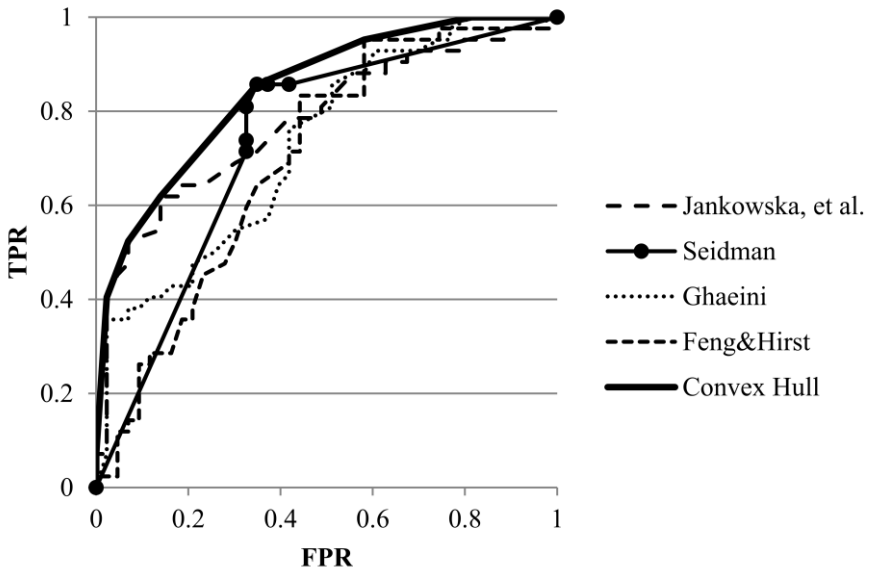

Figure 5: ROC curves of the best performing submissions on the evaluation corpus and their convex hull.

Although the best performance on the English part of the corpus is lower than the best performances on the Greek and Spanish parts, the average performance on the Greek part is lower than the other two parts. Moreover, more submissions are below the baseline for the Greek part of the corpus than the other two parts. We may conclude therefore that the Greek part is more difficult in comparison to the English and the Spanish parts of the corpus.

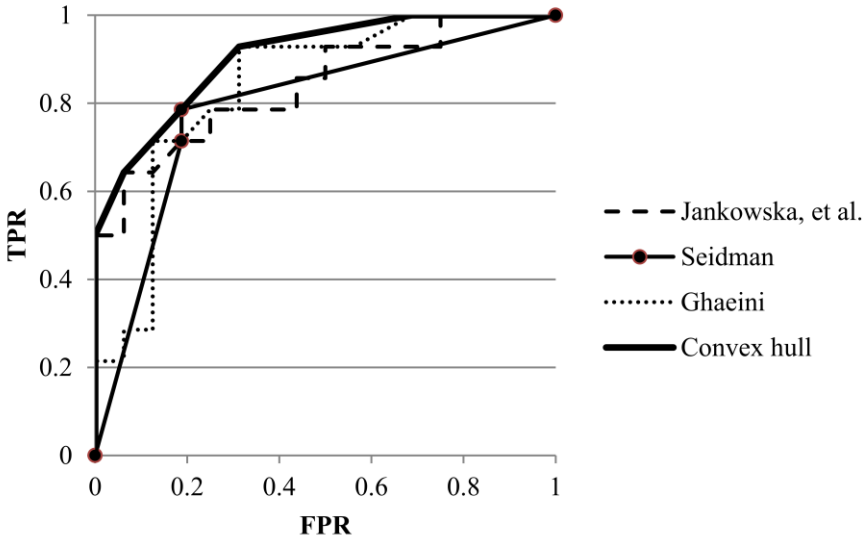

Figure 6: ROC curves of the best performing submissions on the English part of the evaluation corpus and their convex hull.

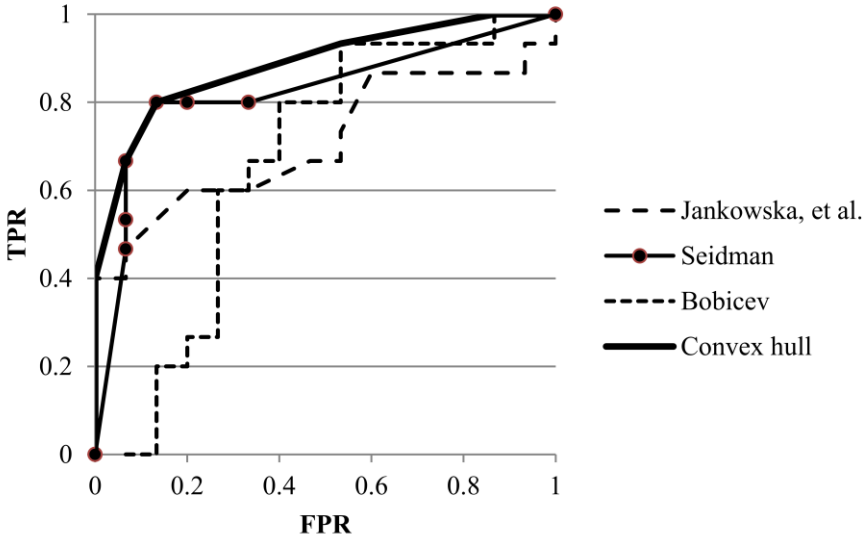

Figure 7: ROC curves of the best performing submissions on the Greek part of the evaluation corpus and their convex hull

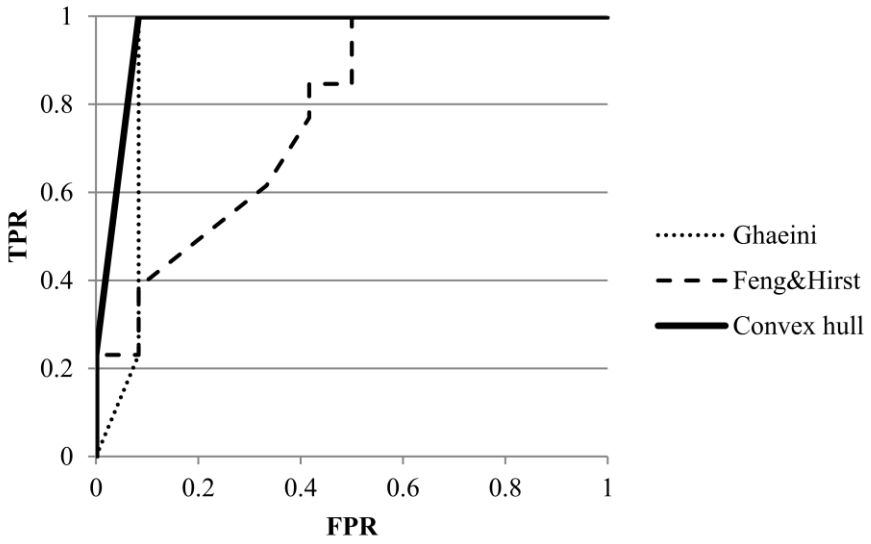

Figure 8: ROC curves of the best performing submissions on the Spanish part of the evaluation corpus and their convex hull

More than half of the participants (i.e., 10 out of 18) have also submitted real scores (i.e., in the set [0,1] inclusive) together with their binary answers. This allowed us to compute the ROC curves and the corresponding AUC values for those participants. The results of this evaluation procedure are shown in Table 5.

The submission of Jankowska *et al*. [13] managed to be equally effective in all three languages. The approach of Seidman [31] was strong in the English and Greek parts but very weak in the Spanish part of the evaluation corpus. The method of Ghaeini [9] was quite remarkable for the Spanish part, strong for the English part but very weak for the Greek part. On the other hand, the submission of Layton *et al*. [25] produces very low AUC scores despite its very good performance using binary answers. A closer examination of the output of their submission indicated that most likely they assign absolute confidence scores in each problem (i.e., assigning 1.0 to a problem they are confident no matter if it is positive or negative) rather than indicating confident positive and confident negative answers as requested.

In more detail, the convex hull of the ROC curves of all the participants on the entire evaluation corpus is depicted in Figure 5. The best performing submissions that form part of the convex hull are also depicted. The corresponding curves per language can be seen in Figures 6, 7, and 8.

The approach of Jankowska *et al*. [13] seems to be more effective for low values of FPR while the approaches of Ghaeini [9], Feng&Hirst [7] and Bobicev [2] work better for high values of FPR. The submission of Seidman [31] seems to be more balanced at least for the English and Greek parts of the corpus. The Spanish part is dominated by the performance of Ghaeini [9].

Table 6. Comparison of early-bird and final evaluation results (F_1).

Submission	Overall	English	Greek	Spanish	Evaluation
Jankowska, et al.	0.720	0.700	0.700	0.800	Early-bird
	0.659	0.733	0.600	0.640	Final
Layton, et al.	0.680	0.750	0.550	0.800	Early-bird
	0.671	0.767	0.500	0.760	Final
Halvani, et al.	0.660	0.750	0.600	0.600	Early-bird
	0.718	0.700	0.633	0.840	Final
Ledesma, et al.	0.620	0.750	0.450	0.700	Early-bird
	0.612	0.467	0.667	0.720	Final
Jayapal&Goswami	0.580	0.600	0.600	0.500	Early-bird
	0.576	0.600	0.633	0.480	Final
Vartapetiance&Gillam	0.560	0.450	0.500	0.900	Early-bird
	0.541	0.500	0.533	0.600	Final
Grozea	0.480	0.450	0.500	0.500	Early-bird
	0.553	0.400	0.600	0.680	Final
Petmanson	0.440	0.500	0.400	0.400	Early-bird
	0.671	0.667	0.567	0.800	Final

Early-bird evaluation: To help participants build their approaches in time we allowed them to submit early versions of their models to be tested using a part of the evaluation corpus. That way, they could identify bugs in their software and fix them and also have an idea of the effectiveness of their models based on real evaluation problems. In total, 8 teams used this option. Table 6 presents the results of the early-bird and final evaluation phases for these teams. Surprisingly, most of the teams participated in the early-bird evaluation phase performed worse in the final evaluation corpus. On the other hand, the submissions of Halvani *et al*. [12] and especially Petmanson [29] took full advantage of this procedure to improve their effectiveness.

Combining the submitted approaches: Having access to the output of all the submitted approaches, we attempted to combine them all into a meta-model. This was inspired by a similar idea applied to the PAN-2010 competition on Wikipedia vandalism detection [30]. Hence, we built a simple meta-classifier based on the binary output of the 18 submitted models. When the majority of the binary answers is Y/N then a positive/negative answer is produced. In ties, a "I don't know" answer is given. Moreover, a real score is generated corresponding to the ratio of the number of positive answers to the number of all the answers. The results of this simple meta-model can be seen in Table 7. By comparing these results with those of the individual submissions, we conclude that the meta-model is in general more effective. It is beaten only by the approach of Seidman [31] for the Greek part of the corpus. As concerns the real confidence scores, again the meta-model is very effective improving the overall performance. However, it is beaten by the approaches of Jankowska *et al*. [13] and Ghaeini [9] in the English part and by Seidman [31] in the Greek part of the corpus. It is remarkable that in the Spanish part the meta-model managed to equal the excellent performance of Ghaeini [9]. In addition, Figure 9 shows the ROC curves of

Table 7. The performance of the meta-classifier combining the output of all the submissions.

	F1	Precision	Recall	AUC
Overall	0.814	0.829	0.800	0.841
English	0.867	0.867	0.867	0.821
Greek	0.690	0.714	0.667	0.756
Spanish	0.898	0.917	0.880	0.926

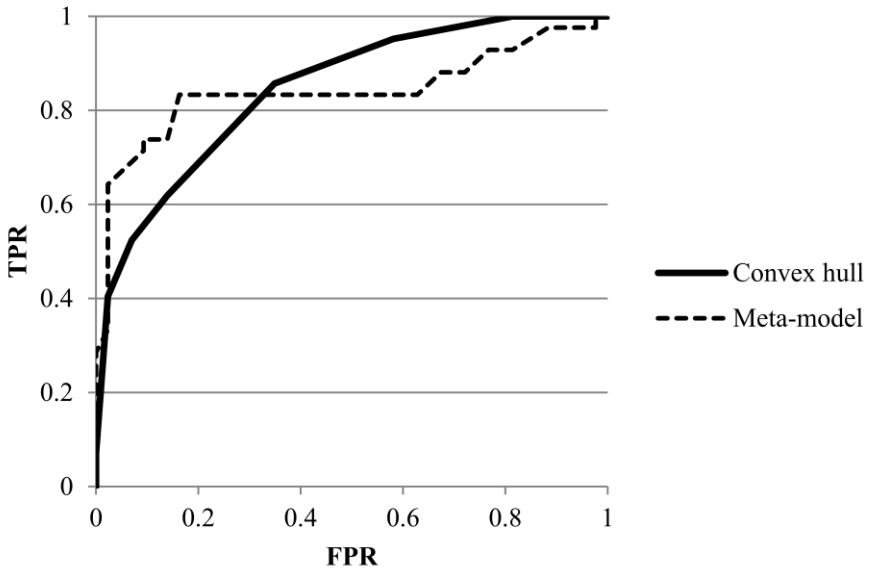

Figure 9. Comparison of the ROC curves of the meta-model and the convex hall of the participants on the entire evaluation corpus.

the meta-model and the convex hull of all the participants based on the entire evaluation corpus. It is clear that the meta-model is more effective for low and medium values of FPR (i.e., it is more accurate in positive answers) while it is weaker for high values of FPR.

7 Survey of the Submitted Approaches

Out of 18 participants, 16 submitted a notebook describing their approach. Here we try to review these approaches. In the following, we use the term *training corpus* to

refer to the collection of verification problems released before the evaluation phase and not the documents of known authorship within a verification problem.

Text representation: The features used by the participants include character, lexical, syntactic, and semantic features [33]. The most popular character features were letter frequencies [7, 12], punctuation mark frequencies [9, 12, 26, 29, 36], character n-grams [5, 12, 14, 25, 27, 31], and common prefixes-suffices of words [12, 19, 36]. Submissions based on compression models are also utilize character sequence patterns [2, 35]. The most widely used lexical features were word frequencies [26, 31], word n-grams [12, 26], function words [7, 9, 12, 36], function word n-grams [12, 34], hapax legomena [7, 19], morphological information (lemma, stem, case, mood, etc.) [9, 29], word, sentence and paragraph length [7, 9, 26], grammatical errors and slang words [19]. Some participants used NLP tools to extract more complex, syntactic and semantic features. POS n-grams are the most popular features of this category [7, 27, 29, 36]. The approach of Vilariño *et al*. [36] build graphs based on POS sequences and then extract sub-graph patterns. Feng&Hirst [7] use POS entropy and more advanced coherence features as discourse-level authorship information by using an NLP tool able to extract entities from and resolve coreferences in English texts. To analyze Greek and Spanish texts, they first translate them to English. In general, the use of NLP tools considerably increases the computational cost [7, 27, 29, 36]. Some participants combine different types of features in their models [9, 12, 36, 27, 29] while others use a single type of features [5, 14, 25, 34]. Similarly, Seidman [31] selects the most appropriate feature type per language.

Classification models: The submitted approaches fall in two main categories: *intrinsic* and *extrinsic* verification models. Intrinsic models are only based on the set of documents of known authorship and the document of unknown authorship to make their decision. Examples of this category are the approaches of Layton *et al*. [29] Halvani *et al*. [12], Jankowska *et al*. [13], and Feng&Hirst [7]. On the other hand, extrinsic models use external resources, that is additional documents by other authors taken from the training corpus or downloaded from the web. Usually extrinsic models attempt to transform the one-class classification problem to a binary or multi-class classification problem. The winning submission [31] follows this approach. The submission of Veenman&Li [35] that is very effective on the English part of the corpus also collects documents of similar genre from the web and builds a two-class classifier. Vilariño *et al*. [36] build a multi-class classifier based on the training corpus and an additional class formed by the documents of known authorship per problem. Moreover, van Dam [5] uses information from the training corpus (i.e., the average distance between the test document and the unknown documents) to decide about a given problem. In addition, the training corpus for English was extended by using additional documents of other authors. In both intrinsic and extrinsic methods, ensemble classification models are very popular and effective [9, 12, 25, 31]. Other popular models are modifications of the CNG method [5, 13, 25], variations of the *unmasking* method [7, 27], and compression-based approaches [2, 35]. The vast majority of the participants follow the *instance-based* paradigm [33] where each document of known authorship is treated separately. In some cases the documents of known authorship are first concatenated and then split into fragments of equal size [2,

12]. Some methods require at least two documents of known authorship, hence in case there is only one such document, they split it into two parts [13, 29]. On the other hand, only the approach of van Dam [5] follows the *profile-based* paradigm where all known documents are treated cumulatively.

Parameter tuning: One basic question is how to optimize the parameter values required by every verification method. In addition, since the evaluation corpus comprises problems in three languages, language-dependent parameter settings should be defined. Some participants avoid this problem by using global parameter settings [9, 12, 14, 26]. However, the majority of the participants used the training corpus sometimes enhanced by external documents found in the web or from other collections to better estimate the appropriate parameter values per language [13, 29, 31]. On the other hand, Layton *et al*. [25] take advantage of this problem by building an ensemble model where each base classifier corresponds to a different configuration of the parameters.

Text normalization: The majority of the approaches did not perform any kind of text preprocessing. They just used the original textual data as found in the set of known documents and the unknown document. Some participants performed simple transformations like the removal of diacritics [5, 12], substitution of digits with a special symbol [5], or conversion of the text to lowercase [5]. More importantly, several participants attempted to normalize the text-length of the documents. Halvani *et al*. [12] and Bobicev [2] first concatenate all known documents and then segment them into equal-size fragments. Jankowska *et al*. [13] reduces all documents within a problem to the same size to produce equal-size representation profiles. This process seems to be crucial especially for methods based on character representations.

8 Discussion

The author identification task at PAN-2013 introduced a number of novelties. First, it required software submissions, therefore enabling reproduction of the results and comparison of runtimes. In addition, the submitted approaches can now easily be applied to any corpus of similar properties and thus it will be possible to be compared with future models. Second novelty is the task definition itself. The problem of having a few documents of known authorship and one document of questioned authorship can model any given author identification task (i.e., multi-class, closed-set, or open-set cases). So, this is a fundamental problem in authorship attribution research [23]. Third, the corpus built in the framework of this task includes verification problems in three natural languages and genres. It tested the ability of the submitted approaches to handle resource-rich languages and resource-poor languages. In addition, the task indirectly posed the question how to appropriately tune a certain method for a given genre/language.

The participation in this task was more than satisfactory. In total, 18 teams from 14 countries have submitted their software. We are aware that certain teams with mainly a linguistic background develop semi-automated approaches to author identification

and therefore had difficulties to submit their methods to this fully-automated evaluation campaign. To enable their participation, we offered an alternative option to such teams so that they have access to the evaluation corpus after the deadline of software submissions and then submit their results to be ranked in a separate list. However, finally there was no such participation. We hope to attract more teams with linguistic background in future evaluation campaigns since our ultimate goal is to provide a common forum for all researchers working on author identification.

The vast majority of the participants answered all the problems of the evaluation corpus. Only two teams used the "I don't know" option. Given the nature of the author verification applications, it is crucial for verification models to only provide the answers they are quasi-certain about. Unfortunately, the performance measures we used in this task do not give enough weight to verification problems left unanswered. In future evaluation campaigns, the performance measures should be better selected towards this direction. For example, the c@1 measure [28] used in the question answering community could be useful. Moreover, the submission of real scores indicating the confidence of the provided answers should be mandatory since ROC curves offer a very detailed picture of the submitted models. Additionally, ROC curves are independent of the distribution of positive/negative problems in the evaluation corpus [6] and therefore the conclusions drawn from this analysis are more general.

The most successful submitted approaches follow the extrinsic verification paradigm where the one-class problem is transformed to a multi-class classification problem, one class formed by the documents of known authorship and the other classes formed by documents of other authors found in external resources [31, 35]. Moreover, methods based on complicated features extracted by specialized NLP tools do not seem to have any advantage over simpler methods based on character and lexical information. The latter require very low computational cost.

The meta-model combining the output of all the submissions proved to be very effective and in average better than any individual method. The combination of heterogeneous models has not attracted much attention so far in authorship attribution research and certainly needs to be examined thoroughly. To this end, it is crucial to increase the publicly-available implementations of certain author identification methods.

It is also important to consider what, if any, changes should be made to future similar evaluations. In our opinion, the same basic verification framework should be retained at least for the next few iterations of PAN/CLEF or similar conferences. This will enable researchers to concentrate their efforts on incremental improvements of the analysis technology itself instead of on meeting changes in the problem specifications. At the same time, in light of the importance of authorship attribution as a forensic problem [4, 11] as well as the emerging need for accuracy standards and "solid linguistic research" into "reliable markers of authorship" [3, 18], it is also important to consider what type of problems to incorporate. Many real-world problems do not have substantial "external resources," whether because they are in less commonly studied languages, historical dialects, or simply unusual genres such as ransom notes. Put simply, what, if any, real-world applications of authorship attribution should be modeled, and how best should the modeling happen? How can PAN frame the problem in order to continue to attract a wide variety of participants,

including not merely computational approaches, but also approaches that use human expertise and high level linguistic information, a feature largely absent from the high scoring participants in this round?

Acknowledgement

This work was partially supported by the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action. This material is based on work supported by the National Science Foundation under grant no. OCI-1032683. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] S. Argamon and P. Juola. Overview of the International Authorship Identification Competition at PAN-2011. In V. Petras, P. Forner, P.D. Clough (eds.) *CLEF Notebook Papers/Labs/Workshop*, 2011.

[2] V. Bobicev. Authorship Detection with PPM – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[3] R.R. Butters. Ethics, Best Practices, and Standards, In *Proc. IAFL'11*, 2011.

[4] M. Coulthard. On Admissible Linguistic Evidence. *J. Law and Policy*, 21:2 441-466, 2013.

[5] M. van Dam. A Basic Character N-gram Approach to Authorship Verification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[6] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8): 861-874, 2006.

[7] V.W. Feng and G. Hirst. Authorship Verification with Entity Coherence and Other Rich Linguistic Features – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[8] P. Forner, R. Navigli, and D. Tufis (eds). *CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers*, 2013.

[9] M.R. Ghaeini. Intrinsic Author Identification Using Modified Weighted KNN – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[10] T. Gollub, B. Stein, S. Burrows, and D. Hoppe. TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In Proc. of TIR at DEXA'12. IEEE.

[11] T. Grant. TXT 4N6: Method, Consistency, and Distinctiveness of the Analysis of SMS Text Messages. *J. Law and Policy*, 21:2 467-494, 2013.

[12] O. Halvani, M. Steinebach, and R. Zimmermann. Authorship Verification via k-Nearest Neighbor Estimation – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[13] M. Jankowska, V. Kešelj, and E. Milios. Proximity based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].

[14] A. Jayapal and B. Goswami. Vector Space Model and Overlap Metric for Author Identification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[15] P. Juola. Ad-hoc Authorship Attribution Competition. In *Proc. of ALLC'04*.
[16] P. Juola. Authorship Attribution. *Foundations and Trends in IR*, 1:234–334, 2008.
[17] P. Juola. An Overview of the Traditional Authorship Attribution Subtask. In *Proc. of CLEF'12*.
[18] P. Juola. A Critical Analysis of the Ceglia/Zuckerberg Email Authorship Study. In *Proc. IAFL'13*, 2013.
[19] R. Kern. Grammar Checker Features for Author Identification and Author Profiling – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[20] M. Koppel and Y. Winter. Determining if Two Documents are by the Same Author. *Journal of the American Society for Information Science and Technology*, to appear.
[21] M. Koppel, J. Schler, and S. Argamon. Computational Methods in Authorship Attribution. *Journal of the American Society for information Science and Technology*, 60(1):9-26, 2009.
[22] M. Koppel, J. Schler, and S. Argamon. Authorship Attribution in the Wild. *Language Resources and Evaluation*, 45:83–94, 2011.
[23] M. Koppel, J. Schler, S. Argamon, and Y. Winter. The "Fundamental Problem" of Authorship Attribution, *English Studies*, 93(3): 284-291, 2012.
[24] M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research*, 8:1261–1276, 2007.
[25] R. Layton, P. Watters, and R. Dazeley. Local n-grams for Author Identification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[26] P. Ledesma, G. Fuentes, G. Jasso, A. Toledo, and I. Meza. Distance Learning for Author Verification – Notebook for PAN at CLEF 2013 In Forner et al. [8]..
[27] E. Moreau and C. Vogel. Style-based Distance Features for Author Verification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[28] A. Peñas and A. Rodrigo. A Simple Measure to Assess Nonresponse. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1415-1424, 2011.
[29] T. Petmanson. Authorship Identification Using Correlations of Frequent Features – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[30] M. Potthast, B. Stein, and T. Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In M. Braschler and D. Harman (eds), *Notebook Papers of CLEF 10 Labs and Workshops*, 2010.
[31] S. Seidman. Authorship Verification Using the Impostors Method – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[32] E. Stamatatos. Author Identification Using Imbalanced and Limited Training Texts, In *Proc. of the 4th International Workshop on Text-based Information Retrieval*, 2007.
[33] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60:538–556, 2009.

[34] A. Vartapetiance ane L. Gillam. A Textual Modus Operandi: Surrey's Simple System for Author Identification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[35] C.J. Veenman and Z. Li. Authorship Verification with Compression Features – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].
[36] D. Vilariño, D. Pinto, H. Gómez, S. León, and E. Castillo. Lexical-Syntactic and Graph-Based Features for Authorship Verification – Notebook for PAN at CLEF 2013. In Forner *et al*. [8].