Author Profiling and Plagiarism Detection

Paolo Rosso

Natural Language Engineering Lab, PRHLT Research Center,
Universitat Politècnica de València, Spain
prosso@dsic.upv.es http://www.dsic.upv.es/prosso

Abstract In this paper we introduce the topics that we will cover in the RuSSIR 2014 course on Author Profiling and Plagiarism Detection (APPD). Author profiling distinguishes between classes of authors studying how language is shared by classes of people. This task helps in identifying profiling aspects such as gender, age, native language, or even personality type. In case of the plagiarism detection task we are not interested in studying how language is shared. On the contrary, given a document we are interested in investigating if the writing style changes in order to unveil text inconsistencies, i.e., unexpected irregularities through the document such as changes in vocabulary, style and text complexity. In fact, when it is not possible to retrieve the source document(s) where plagiarism has been committed from, the intrinsic analysis of the suspicious document is the only way to find evidence of plagiarism. The difficulty in retrieving the source of plagiarism could be due to the fact that the documents are not available on the web or the plagiarised text fragments were obfuscated via paraphrasing or translation (in case the source document was in another language). In this overview, we also discuss the results of the shared tasks on author profiling (gender and age identification) and plagiarism detection that we help to organise at the PAN Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse (http://pan.webis.de).

1 Author Profiling: How Writing Style is Shared

Author profiling tries to determine an author's gender, age, native language, personality type, etc. solely by analysing her texts. Profiling anonymous authors is a problem of growing importance, both from forensic and marketing perspectives. From a forensic perspective it is important to identify the linguistic profile of an author of a harassing text message or a potential online paedophile on the basis of the analysis of his writing style in order, for instance, to unveil his age [58] [7]. From a marketing viewpoint, companies may be interested in knowing the demographics of their target group in order to achieve a better market segmentation.

In this section we will introduce the reader to the profiling aspects of gender and age identification, describing the shared task that was organised at PAN, and briefly discussing the obtained results and the way the problem was addressed by the participants. PAN was the first lab to offer author profiling as a shared task. At PAN 2013 [58] we aimed at identifying age and gender from a large corpus collected from social media. Most of the participants used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, and so on, together with POS tags and content-based features such as latent semantic analysis, bag-of-words, tfidf, dictionary-based words, topic-based words, and so on. The winner of the PAN 2013 task [29] used second order representations based on relationships between documents and profiles, whereas another well-performing approach, winner of the English subtask [35], used collocations. Following we summarise the evaluation of 10 author profilers that have been submitted to the shared task that was organised in 2014.

Evaluation Corpora In the author profiling task at PAN 2013 [58] participants approached the task of identifying age and gender in a large corpus collected from social media, and age was annotated with three classes: 10s (13-17), 20s (23-27), and 30s (33-47). At PAN 2014, we continued to study the gender and age aspects of the author profiling problem, however, four corpora of different genres were considered—social media, blogs, Twitter, and hotel reviews—both in English and Spanish. Moreover, we annotated age with the following continuous classes: 18-24; 25-34; 35-49; 50-64; and 65+.

The social media corpus was built by sampling parts of the PAN 2013 evaluation corpus. We selected only authors with an average number of words greater than 100 in their posts. We also reviewed manually the data in order to remove authors who appeared to be fake profiles such as bots. The blogs and Twitter corpora were manually collected and annotated by three annotators. The Twitter corpus was built in collaboration with RepLab,[1] where the main goal of author profiling in the context of reputation management on Twitter was to decide how influential a given user is in a domain of interest. For each blog, we provided up to 25 posts and for each twitter profile, we provided up to 1000 tweets. The hotel review corpus is derived from another corpus that was originally used for aspect-level rating prediction [69].[2] The original corpus was crawled from the hotel review site TripAdvisor[3] and manually checked for quality and compliance with the format requirements of PAN 2014.

Evaluation Results In Table 1 joint identification accuracies for both gender and age prediction are shown per data set and averaged over all corpora, which also serves as ranking criterion. The baseline considered the 1000 most frequent character trigrams. In summary, simple content features, such as bag-of-words or word n-grams achieve best accuracies. Bag-of-words features are used by Liau and Vrizlynn [28], word n-grams are used by Maharjan *et al.* [31], and term vector models are used by Villena-Román and González-Cristóbal [67]. They achieved competitive performances on almost all corpora. Weren *et al.* [70] employed information retrieval features and Marquardt *et al.* [32] mixed content and style features. Some readability measures were also used: Automated Read-

[1] http://nlp.uned.es/replab2014

[2] http://times.cs.uiuc.edu/˜wang296/data

[3] http://www.tripadvisor.com

Table 1. Author profiling: joint identification (gender and age) results in terms of accuracy.

Team	Overall	Social Media		Blogs		Twitter		Reviews
		en	es	en	es	en	es	en
López-Monroy	**0.2895**	0.1902	0.2809	**0.3077**	**0.3214**	**0.3571**	0.3444	0.2247
Liau	0.2802	0.1952	**0.3357**	0.2692	0.2321	0.3506	0.3222	**0.2564**
Shrestha	0.2760	**0.2062**	0.2845	0.2308	0.2500	0.3052	**0.4333**	0.2223
Weren	0.2349	0.1914	0.2792	0.2949	0.1786	0.2013	0.2778	0.2211
Villena-Román	0.2315	0.1905	0.1961	**0.3077**	0.2321	0.2078	0.2667	0.2199
Marquardt	0.1998	0.1428	0.2102	0.1282	0.2679	0.1948	0.3111	0.1437
Baker	0.1677	0.1277	0.1678	0.1282	0.2321	0.1688	0.2111	0.1382
Baseline	*0.1404*	*0.0930*	*0.1820*	*0.0897*	*0.0536*	*0.1494*	*0.2333*	*0.1821*
Mechti	0.1067	0.1244	0.1060	0.0897	0.1786	0.0584	0.1444	0.0451
Castillo Juarez	0.0946	0.1445	0.1254	0.1795	0.0893	–	–	0.1236
Ashok	0.0834	0.1318	–	0.1282	–	0.1948	–	0.1291

ability index [32, 20], Coleman-Liau index [32, 20], Rix Readability index [32, 20], Gunning Fog index [20], Flesch-Kinkaid [70]. The approach of López-Monroy *et al.* [30] obtained the best overall using a second order representation based on relationships between documents and profiles.

From the results of Table 1, it can be seen that: *a*) the highest joint accuracies were achieved on Twitter data, and, *b*) the smallest joint accuracies were achieved in English social media and hotel reviews. It is an open question why these differences can be observed, whereas possible explanations may be that people express themselves more spontaneously on Twitter compared to the other genres, whereas the low scores are due to the approaches' difficulty of predicting gender in social media and age in hotel reviews. A complete version of the report can be found in [59], where a more in-depth analysis of the obtained results as well as a survey of detection approaches are given.

2 Plagiarism Detection: How Writing Style Changes

Plagiarism is the re-use of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source [26]. A person that fails to provide its corresponding source is suspected of plagiarism. In the academic domain, some surveys estimate that around 30% of student reports include plagiarism [2] and a more recent study increases this percentage to more than 40% [10]. Indeed the amount of text available in electronic media nowadays has caused cases of plagiarism to increase. As a result, its manual detection has become infeasible. Models for automatic plagiarism detection are being developed as a countermeasure. Their main objective is assisting people in the task of detecting plagiarism—as a side effect, plagiarism is discouraged.

However, not always it is straightforward to retrieve the document(s) that have been the source of plagiarism because they may be not available or with a high level of paraphrasing or even in another language. In this section we describe basic stylistic analysis techniques in order to spot irregularities in the writing style of the suspicious document. As well we illustrate how performance of plagiarism detectors decreases in case of obfuscation of the source text via paraphrasing or translation.

2.1 Stylistic Analysis

When it is not possible to retrieve the document(s) plagiarism has been committed from, or because they are not available in the given collection (or even on the web) or due to the high level of obfuscation via paraphrasing or translation, the evidence of plagiarism has to be found in the document itself (intrinsic plagiarism detection). The aim is to spot changes in vocabulary, text complexity and writing style (see Figure 1). In fact, the insertion of text fragments from a different author into the suspicious document causes style and complexity irregularities. The quantification can be made by measuring vocabulary richness (type/tokens ratio), basic statistics (average sentence length, average word length, etc.), n-gram profiles (character level statistics), and text readability measures (e.g. Gunning Fog, Flesch-Kinkaid, etc.) [36, 62]. We have already mentioned in the previous section how readability measures help in author profiling as well. In fact, complexity in texts and writing style change with the author's demographics (e.g. her age, gender or personality). The formula below refers to the Gunning Fog (GF) index which, in order to determine the complexity of a given text, takes into account its total number of sentences, words and complex words, where complex words are those words with three or more syllables [23]. The value resulting of this calculation can be interpreted as the number of years of formal education required to understand the document contents.

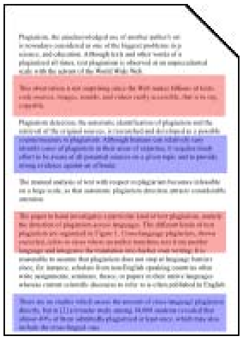


Figure 1. Identifying changes in writing style within a document.

$$GF = 0.4(\frac{|words|}{|sentences|} + 100 * \frac{|complex - words|}{|words|}) \tag{1}$$

Typical values for GF of texts such as a comic, a Newsweek article, and scientific texts are: GF(comic)=6, GF(Newsweek)=10, GF(T1)=15.2, and GF(T2)=14.1

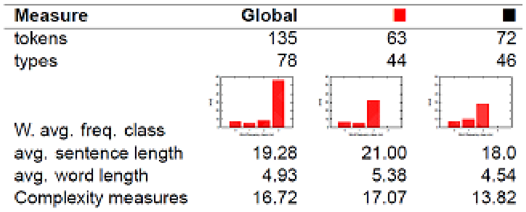
Let us analyse the three text fragments of the example below.

Example 1. In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. [...]

Mineral salts are inorganic molecules of easy ionization in presence of water; in living beings they appear by precipitation as well as dissolved mineral salts. [...] Dissolved mineral salts are always ionized. These salts have a structural function and pH regulating functions, of the osmotic pressure and and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. [...]

Figure 2 illustrates statistics and measures used for stylistic analysis. Values for the first and third paragraphs (third column) are in back, whereas the values for the more formal second paragraph (second column) are in red.



Measure	Global	🟥	⬛
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.0
avg. word length	4.93	5.38	4.54
Complexity measures	16.72	17.07	13.82

Figure 2. Stylisic analysis.

Although when the suspicious document is short an expert simply reading it could quite easily detect text inconsistencies and writing style changes, when the document is long (e.g. a thesis or a report) spotting irregularities in text is not always straightforward. Therefore, it is important to have tools that could help the experts (e.g. forensic linguists, teachers, etc.) in highlighting suspicious text fragments. For instance, Stylysis[4] is a tool whose aim is to provide the expert with a linguistic profile of the document on the basis of a stylistic analysis in order to determine whether or not there are text fragments of different writing styles. Stylysis analyses documents in English, Spanish or Catalan. The tool divides the text into fragments and for each of them, it calculates basic statistics, as well as vocabulary richness measure (function K proposed by Yule [72] and function R proposed by Honore [25]) and text readability measures (Gunning Fog index [23] and Flesch-Kincaid Readability test [11]).

[4] http://memex2.dsic.upv.es/StylisticAnalysis

2.2 Obfuscation via Paraphrasing

The linguistic phenomena underlying plagiarism have barely been analyzed. In [33] different kinds of plagiarism are identified: of ideas, of references, of authorship, word by word, and paraphrase plagiarism. In the first case, ideas, knowledge or theories from another person are claimed without proper citation. In plagiarism of references and authorship, citations and entire documents are included without any mention of their authors. Word by word plagiarism, also known as copy–paste or verbatim copy, consists of the exact copy of a text (fragment) from a source into the plagiarised document. Regarding paraphrase plagiarism, a different form expressing the same content is used.

For this purpose we will show how the plagiarism detectors that participated in the PAN shared task in 2010 decreased their performance on the subset of paraphrase plagiarism cases of the PAN-PC-10 corpus [48]. First, we briefly describe the evaluation measures that are employed in the shared task on plagiarism detection [45].

Evaluation Measures As automatic plagiarism detection is identified as an information retrieval task, evaluation is usually carried out on the basis of recall and precision. Nevertheless, plagiarism detection aims at retrieving plagiarised–source fragments rather than documents. Given a suspicious document d_q and a collection of potential source documents D, the detector should retrieve: *a*) a specific text fragment $s_q \in d_q$, potential case of plagiarism; and *b*) a specific text fragment $s \in d$, the claimed source for s_q. Therefore, special versions of precision and recall have been proposed at PAN in order to fit in this framework. The plagiarized text fragments are treated as basic retrieval units, with $s_i \in S$ defining a query for which a plagiarism detection algorithm returns a result set $R_i \subseteq R$. The recall and precision of a plagiarism detection algorithm are defined as:

$$prec_{PDA}(S,R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|} \quad \text{and} \tag{2}$$

$$recall_{PDA}(S,R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|} \tag{3}$$

where $\sqcap$ computes the positionally overlapping characters. In both equations, S and R represent the entire set of actually plagiarized text fragments and detections, respectively.

Consider Fig. 3 for an illustrative example. $\{s_1, s_2, s_3\} \in S$ represent text sequences in the document that are known to be plagiarised. A given detector recognises the sequences $\{r_1, r_2, r_3, r_4, r_5\} \in R$ as plagiarised. Substituting the values in Equations 2 and 3:

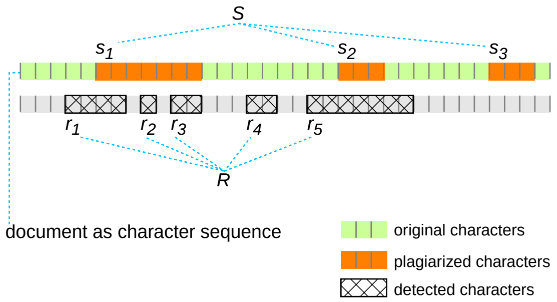


Figure 3. A document as character sequence, including plagiarized sections S and detections R returned by a plagiarism detector.

$$
\begin{aligned}
precision_{PDA}(S,R) &= \frac{1}{|R|}\cdot\left(\frac{|r_1 \sqcap s_1|}{|r_1|}+\frac{|r_2 \sqcap s_1|}{|r_2|}+\frac{|r_3 \sqcap s_1|}{|r_3|}+\cancelto{0}{\frac{|\emptyset|}{|r_4|}}+\frac{|r_5 \sqcap s_2|}{|r_5|}\right)\\
&= \frac{1}{5}\cdot\left(\frac{2}{4}+\frac{1}{1}+\frac{2}{2}+\frac{3}{7}\right) = 0.5857 \quad \text{and}
\end{aligned}
$$

$$
\begin{aligned}
recall_{PDA}(S,R) &= \frac{1}{|S|}\cdot\left(\frac{|(s_1 \sqcap r_1)\bigcup(s_1 \sqcap r_2)\bigcup(s_1 \sqcap r_3)|}{|s_1|}+\frac{|s_2 \sqcap r_5|}{|s_2|}+\cancelto{0}{\frac{\emptyset}{|s_3|}}\right)\\
&= \frac{1}{3}\cdot\left(\frac{5}{7}+\frac{3}{3}\right) = 0.5714
\end{aligned}
$$

Once precision and recall are computed, they are combined into their harmonic mean (F_1-measure).

Evaluation Results Figure 4 (a) shows the evaluations computed by considering the entire PAN-PC-10 corpus. The best recall values are around 0.70, with very good values of precision, some of them above 0.90. The results, when considering only the simulated cases, that is, those generated by manual paraphrasing, are presented in Fig. 4 (b) . In most of the cases, the quality of the detections decreases dramatically compared to the results on the entire corpus, which also contains translated, verbatim and automatically modified plagiarism. The difficulty to detect paraphrase plagiarism cases in the PAN-PC-10 corpus was also stressed in [64]. Manually created cases seem to be much harder to detect than the other, artificially generated, cases (when the simulated cases in the PAN-PC-10 corpus were generated, volunteers had specific instructions to create rewritings with a high obfuscation degree). This can be appreciated when looking at the difference of capabilities of the plagiarism detector applied at the 2009 and 2010 shared tasks by [22] and [21], practically the same implementation. At the first shared task, whose corpus included artificial cases only, its recall was of 0.66 while in the second one, with simulated (i.e., paraphrase plagiarism) cases, it decreased to 0.48.

Interestingly, the best performing plagiarism detectors on the paraphrase plagiarism corpus are not the ones that performed the best at the PAN-10 shared task. For instance, this is the case of [39] that apply greedy string tiling, which aims at detecting as long as possible identical text fragments. This approach outperforms the rest of detectors when dealing with cases with a high density of identical fragments (with paraphrase plagiarism cases in between).

The complete analysis of the results can be found in [6], where a paraphrase typology is employed in order to investigate further the relationship between paraphrasing and plagiarism. Figure 4 (*c*) shows the evaluation results when considering only the cases included in the P4Psubset of the corpus that was annotated with the types of the paraphrase typology.[5]

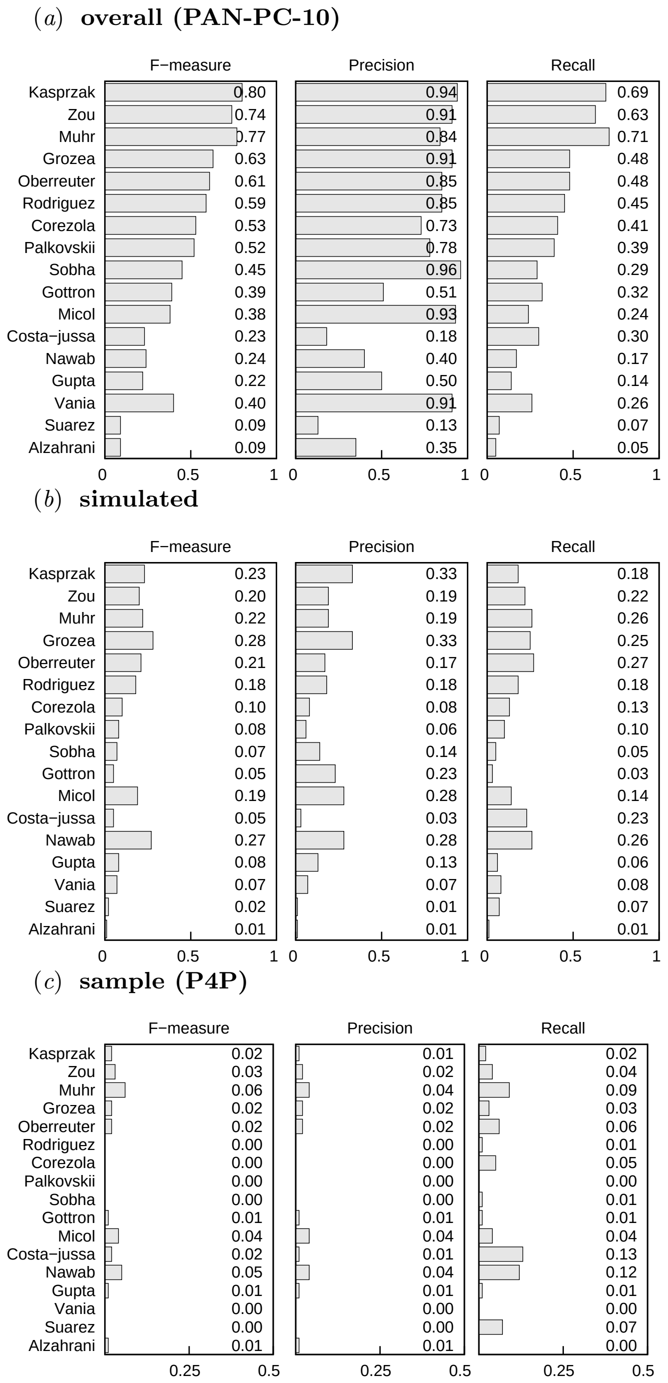
2.3 Obfuscation via Translation

The detection of plagiarism is even more difficult when it concerns documents written in different languages. Cross-language (CL) plagiarism detection attempts to identify and extract automatically plagiarism among documents in different languages. Recently a survey was done on scholar practices and attitudes [3], also from a cross-language plagiarism perspective which manifests that CL plagiarism is a real problem: in fact, only 36.25% of students think that translating a text fragment and including it into their report is plagiarism.

In recent years there have been a few approaches to cross-language similarity analysis that can be used for CL plagiarism detection [5]. A simple, yet effective approach is the cross-language character n-gram (CL-CNG) model [34]. Using character n-grams, it takes into account the syntax of documents, and offers remarkable performance for languages with syntactic similarities. Cross-language explicit semantic analysis (CL-ESA) [46, 50] represents a document by its similarities to a collection of documents. These similarities in turn are computed with a monolingual retrieval model such as the vector space model. The cross-language alignment-based similarity analysis (CL-ASA) model [4, 50] is instead based on statistical machine translation and combines probabilistic translations, using a statistical bilingual dictionary and similarity analysis. The cross-language conceptual thesaurus based similarity (CL-CTS) model [24] tries to measure the similarity between the documents in terms of shared concepts, using a conceptual thesaurus, and named entities among them.

Plagiarised fragments can be translated verbatim copies or may alter their structure to hide the copying, which is known as paraphrasing and is more difficult to detect. In order to improve the detection of paraphrase plagiarism, a model named cross-language knowledge graph analysis (CL-KBS) was introduced [15]. Its goal is to exploit explicit semantics for a better representation of the documents. CL-KGA provides a context model by generating knowledge

[5] http://clic.ub.edu/corpus/en/paraphrases-en



(*a*) **overall (PAN-PC-10)**

(*b*) **simulated**

(*c*) **sample (P4P)**

Figure 4. Evaluation of the PAN-10 competition participants' plagiarism detectors. Figures show evaluations over: (*a*) entire PAN-PC-10 corpus (including artificial, translated, and simulated cases); (*b*) simulated cases only; (*c*) sample of simulated cases annotated on the basis of the paraphrases typology: the P4P corpus. Note the change of scale in (*c*) .

graphs that expand and relate the original concepts from suspicious and source text fragments. Finally, the similarity is measured in a semantic graph space. In this section we compare CL-KGA with CL-ASA and CL-CNG because obtaining the best results in a previous study [50].

Cross-Language Character N-Grams The cross-language character n-gram (CL-CNG) model has shown to improve the performance of CL information retrieval for syntactically similar languages. This model typically uses character trigrams (CL-C3G) to compare documents in different languages [50].

Given a source document d written in a language L_1 and a suspicious document d' written in language L_2, the similarity $S(d, d')$ between the two documents is measured as follows:

$$S(d, d') = \frac{\boldsymbol{d} \cdot \boldsymbol{d'}}{|d| \cdot |d'|}, \tag{4}$$

where $\boldsymbol{d}$ and $\boldsymbol{d'}$ are the vector representation of documents d and d' into character n-gram space.

Cross-Language Alignment based Similarity Analysis The cross-language alignment based similarity analysis (CL-ASA) model measures the similarity between two documents d and d', from two different languages L_1 and L_2 respectively, by aligning the documents at word level and determining the probability of d' being a translation of d. The similarity $S(d, d')$ between both documents is measured as in Equation 5:

$$S(d, d') = l(d, d') * t(d|d'), \tag{5}$$

where $l(d, d')$ is the length factor defined in [56], which is used as normalization since two documents with the same content, in different languages do not have the same length. Moreover, $t(d|d')$ is the translation model defined in Equation 6:

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x, y), \tag{6}$$

where $p(x, y)$ is the probability of a word x from language L_1 being a translation of word y from L_2. These probabilities can be obtained using a bilingual statistical dictionary.

Cross-Language Knowledge Graph Analysis The cross-language knowledge graphs analysis (CL-KGA) model uses knowledge graphs generated from a multilingual semantic network (MSN) in order to obtain a context model of text fragments in different languages. We employ BabelNet [38], although the graph-based model is generic and could be applied with other available MSNs such as EuroWordNet [68].

A knowledge graph is a weighted and labelled graph that expands and relates the original concepts of a set of words, providing a “context model”. Using BabelNet to build the graphs we can have a multilingual dimension for each of the concepts. Therefore, we can compare directly pairs of graphs built from text fragments in different languages to detect CL plagiarism.

We can build a knowledge graph using a MSN such as BabelNet as follows: having a concept set C, we search BabelNet for paths connecting each pair $c, c‘ \in C$, obtaining the set of paths P, where each $p \in P$ is a set of concepts and relations between concepts from C which include the conceptual expansion. The knowledge graph g is obtained after joining the paths from P including all its concepts and relations. Finally, to weight the concepts we use their degree of relatedness, i.e. the number of outgoing edges for each node. The relation weighting is performed also in function of the degree of relatedness of their source and target concepts.

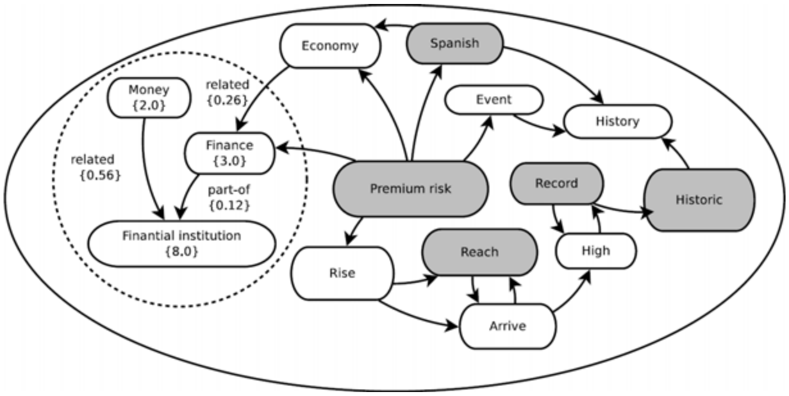


Figure 5. Knowledge graph built from the sentence "Spanish premium risk reaches historic records", simplified without the multilingual dimension, and with labels and weights only inside the dashed circle.

Example 3. Having the English sentence of Example 2, we obtain its concepts $C = \{$Spanish, premium risk, reach, record, historic$\}$. Using BabelNet to build a knowledge graph g from C, we obtain a concept set $C_g = C \cup C'$, where $C' = \{$economy, finance, history...$\}$ is the expanded concept set. In addition, we obtain a relation set $R \in \{$related-to, has-part, belong-to, is-a...$\}$ between concepts of C_g. We can observe the resulting graph g in Fig. 5.

To compare graphs we use a similarity function S that is an adapted version of flexible comparison of conceptual graphs similarity algorithm presented in [37].

$$S(g, g') = S_c(g, g') * (a + b * S_r(g, g')) \tag{7}$$

$$S_c(g,g') = \frac{\left(2 * \sum_{c \in g_{int}} w(c)\right)}{\left(\sum_{c \in g} w(c) + \sum_{c \in g'} w(c)\right)} \tag{8}$$

$$S_r(g,g') = \frac{\left(2 * \sum_{r \in N(c,g_{int})} w(r)\right)}{\left(\sum_{r \in N(c,g)} w(r) + \sum_{r \in N(c,g')} w(r)\right)} \tag{9}$$

where S_c is the score of the concepts, S_r is the score of the relations, a and b are smoothing variables to give the appropriate relevance to concepts and relations, c is a concept, r is a relation, g_{int} is the resulting graph of the intersection between g and g', and $N(c,g)$ is the set of all the relations connected to the concept c in a given graph g.

Evaluation Corpus and Measures In our evaluation we use the German-English (DE-EN) and Spanish-English (ES-EN) CL plagiarism partitions of the PAN-PC'11 corpus [51]. We evaluate the performance of CL-KGA differentiating plagiarism cases between translated verbatim copies and paraphrase translations in which their structure was changed in order to hide the copying [51]. We compare the results obtained by CL-KGA with those provided by CL-ASA and CL-C3G (CL-CNG using 3-grams) for the same task. For CL-ASA model we use two statistical dictionaries: BabelNet's statistical dictionary (CL-ASA$_{BN}$ [15]) and a dictionary trained using the word-aligment model IBM M1 [42] on the JRC-Acquis [65] corpus.

With respect to the evaluation measures, apart from precision and recall at character level, the granularity measure was considered. In fact, due that neither precision nor recall account that plagiarism detectors sometimes report overlapping or multiple detections for a single plagiarism case, the granularity measure has been introduced in the PAN shared task:

$$granularity(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|, \tag{10}$$

where $S_R \subseteq S$ are cases detected by detections in R, and $R_s \subseteq R$ are detections of s; i.e., $S_R = \{s \mid s \in S \text{ and } \exists r \in R : r \text{ detects } s\}$ and $R_s = \{r \mid r \in R \text{ and } r \text{ detects } s\}$.

Finally, the three measures are combined into a single overall score to allow for a unique ranking among detection approaches [45]. :

$$PlagDet(S,R) = \frac{F_1}{\log_2(1 + granularity(S,R))}, \tag{11}$$

Table 2. DE-EN cross-language plagiarism detection results for automatic and paraphrase translation cases, displayed in the decreasing order of the PlagDet score.

Model	German-English							
	Automatic translations				Paraphrase translations			
	PlagDet	Recall	Precision	Granularity	PlagDet	Recall	Precision	Granularity
CL-KGA	**0.5296**	**0.4671**	**0.6306**	**1.0188**	**0.1006**	**0.2101**	**0.0661**	**1.0**
CL-ASA$_{IBMM1}$	0.4230	0.3690	0.6019	1.1163	0.0462	0.0978	0.0303	1.0
CL-ASA$_{BN}$	0.3019	0.2363	0.5962	1.1753	0.0275	0.0796	0.0166	1.0
CL-C3G	0.0909	0.0564	0.3414	1.0913	0.0185	0.0389	0.0121	1.0

Evaluation Results As we can see in Table 2, for the DE-EN CL plagiarism detection, CL-C3G obtains the lowest results, being the baseline for this kind of experiments, due to the simplicity of the approach which uses n-grams. CL-ASA$_{BN}$ uses BabelNet's statistical dictionary, obtaining average results, despite many German words in the dictionary were not found. CL-ASA$_{IBMM1}$ outperforms the baseline *PlagDet* by 365% in automatic translations and 149% in paraphrase translations. Finally, CL-KGA obtains the best values, increasing the baseline *PlagDet* by 478% in automatic translations and 443% in paraphrase translations, along with better values for recall, precision and granularity.

Table 3. ES-EN cross-language plagiarism detection results for automatic and paraphrase translation cases, displayed in the decreasing order of the PlagDet score.

Model	Spanish-English							
	Automatic translations				Paraphrase translations			
	PlagDet	Recall	Precision	Granularity	PlagDet	Recall	Precision	Granularity
CL-KGA	**0.6087**	**0.5399**	**0.7036**	**1.0050**	**0.0993**	**0.1979**	**0.0662**	**1.0**
CL-ASA$_{BN}$	0.5793	0.5245	0.6631	1.0154	0.0738	0.1909	0.0457	1.0
CL-ASA$_{IBMM1}$	0.5339	0.4728	0.6911	1.0729	0.0612	0.1501	0.0384	1.0
CL-C3G	0.1756	0.1336	0.6158	1.3796	0.0289	0.0587	0.0192	1.0

As we can see in Table 3, for ES-EN CL plagiarism detection, the models performance was quite similar to the one obtained for DE-EN. CL-C3G is the baseline with the lowest values. CL-ASA$_{BN}$ increases the baseline *PlagDet* by 230% in automatic translations and 155% in paraphrase translations. This time CL-ASA$_{BN}$ obtains better results than CL-ASA$_{IBMM1}$ showing that using BabelNet's statistical dictionary for ES-EN plagiarism detection allows to obtain a good performance. CL-KGA obtains the best values with all the measures, increasing the baseline *PlagDet* by 246% in automatic translations and 243% in paraphrase translations. The *granularity* for CL-KGA is the closest to 1.0, the best possible value. A more detailed analysis can be found in [16].

3 Related Work

3.1 Author Profiling

The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as psychology, linguistics and, more recently, computational linguistics. In the first section we already mentioned how the teams that participated in the PAN shared task approached author profiling. In this section we describe some of the previous works.

Argamon *et al.* [1] analysed formal written texts extracted from the British National Corpus, combining function words with part-of-speech features and achieving approximately 80% accuracy in gender prediction. Koppel *et al.* [27] studied the problem of automatically determining an author's gender in social media by proposing combinations of simple lexical and syntactic features, and achieving approximately 80% accuracy. Schler *et al.* [61] studied the effect of age and gender in the writing style in blogs; they gathered over 71,000 blogs and obtained a set of stylistic features like non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. They modeled age in three classes —10s (13-17), 20s (23-27) and 30s (33-47)— obtaining an accuracy of about 80% for gender identification and about 75% for age identification. They showed that language features in blogs correlate with age, as reflected in, for example, the use of prepositions and determiners. Goswami *et al.* [19] added some new features as slang words and the average length of sentences, improving accuracy to 80.3% in age group identification and to 89.2% in gender detection. More recently, Nguyen *et al.* [40] studied the use of language and age among Dutch Twitter users. They modelled age as a continuous variable and used an approach based on logistic regression. They measured the effect of the gender in the performance of age identification, considering both variables as inter-dependent, and achieved correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years. Pennebaker *et al.* [44] connected language use with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding the gender and age of its author.

A shared task on computational personality recognition was recently organised at the WCPR workshop of ICWSM 2013[6] and at ACM Multimedia 2014.[7] Moreover, a shared task was also organised at the BEA-8 Workshop of NAACL-HLT 2013 on another aspect of author profiling: native language identification.[8] The number of shared tasks on different aspects of author profiling (gender and age identification, personality recognition, and native language identification) show the raising interest of the scientific community in this challenging problem.

3.2 Plagiarism Detection

In recent years, the evaluation of plagiarism detectors has been studied in the context of the PAN evaluation labs that have been organised annually since 2009.[9] During the first three labs, a total of 43 plagiarism detectors have been evaluated using this framework [47, 49, 51]. The two recent editions re-focused on specific sub-problems of plagiarism detection: source retrieval and

[6] http://mypersonality.org/wiki/doku.php?id=wcpr13

[7] https://sites.google.com/site/wcprst/home/wcpr14

[8] https://sites.google.com/site/nlisharedtask2013/

[9] The corpora PAN-PC-2009/2010/2011 are available at http://www.webis.de/research/corpora

Table 4. Plagiarism detection: source retrieval results.

Team (alphabetical order)	**Downloaded Sources**			**Total Workload**		**Workload to 1st Detection**		**No Detect.**	**Runtime**
	F_1	precision	recall	Queries	Dwlds	Queries	Dwlds		
Elizalde	0.34	0.40	0.39	54.5	33.2	16.4	3.9	7	**04:02:00**
Kong	0.12	0.08	0.48	83.5	207.1	85.7	24.9	6	24:03:31
Prakash	0.39	0.38	**0.51**	60.0	38.8	8.1	3.8	7	19:47:45
Suchomel	0.11	0.08	0.40	**19.5**	237.3	**3.1**	38.6	**2**	45:42:06
Williams	**0.47**	**0.57**	0.48	117.1	**14.4**	18.8	**2.3**	4	39:44:11
Zubarev	0.45	0.54	0.45	37.0	18.6	5.4	**2.3**	3	40:42:18

text alignment. Both source retrieval and text alignment have been identified as integral parts of plagiarism detection [63]. Instead of again applying a semiautomatic approach to corpus construction, a large corpus of manually generated plagiarism has been crowdsourced in order to increase the level of realism [18]. This corpus comprises 297 essays of about 5000 words length, written by professional writers. In this regard the writers were given a set of topics to choose from along with two more technical rules: *i*) to use the ChatNoir search engine [53] to research their topic of choice, and *ii*) to reuse text passages from retrieved web pages in order to compose their essay. The resulting essays represent the to-date largest corpus of realistic text reuse cases available, and they have been employed to evaluate another 33 plagiarism detectors in the past three labs [52, 54, 55].

Source Retrieval In source retrieval, given a suspicious document and a web search engine, the task is to retrieve all source documents from which text has been reused whilst minimizing retrieval costs. To study this task, we employ a controlled, static web environment, which consists of a large web crawl and search engines indexing it. Using this setup, we built the previously described large corpus of manually generated text reuse in the form of essays, which serve as suspicious documents and which are fed into a plagiarism detector.

Table 4 shows the performances of the six plagiarism detectors that implemented source retrieval. Their cost-effectiveness is measured as average workload per suspicious document, and as average numbers of queries and downloads until the first true positive detection has been made. These statistics reveal if a source retrieval algorithm finds sources quickly, thus reducing its usage costs. Moreover, we measure precision and recall of downloaded documents with regard to the true source documents and compute F_1.

None of the detectors dominates the others in terms of all of the employed measures, whereas three detectors share the top scores among them. The detector of Williams *et al.* [71] achieves the best trade-off between precision and recall in terms of F_1 as well as best precision, whereas the detector of Prakash and Saha [57] achieves best recall. Suchomel and Brandejs [66]'s detector requires least query workload, least queries until first detection, and detects source documents for almost all of the test documents. The detector of Williams *et al.* [71], however, performs worst in terms of total querying workload, since it requires 117 queries on average. Posing a query to a search engine may entail significant costs, whereas downloading a document is considered much less costly. By comparison, the detector of Zubarev and Sochenkov [73] achieves a similarly good

Table 5. Plagiarism detection: text alignment results.

Team	PlagDet	Recall	Precision	Granularity	Runtime
Sanchez-Perez	**0.87818**	**0.87904**	0.88168	1.00344	00:25:35
Oberreuter	0.86933	0.85779	0.88595	1.00369	00:05:31
Palkovskii	0.86806	0.82637	0.92227	1.00580	01:10:04
Glinos	0.85930	0.79331	**0.96253**	1.01695	00:23:13
Shrestha	0.84404	0.83782	0.85906	1.00701	69:51:15
R. Torrejón	0.82952	0.76903	0.90427	1.00278	**00:00:42**
Gross	0.82642	0.76622	0.93272	1.02514	00:03:00
Kong	0.82161	0.80746	0.84006	1.00309	00:05:26
Abnar	0.67220	0.61163	0.77330	1.02245	01:27:00
Alvi	0.65954	0.55068	0.93375	1.07111	00:04:57
Baseline	*0.42191*	*0.34223*	*0.92939*	*1.27473*	*00:30:30*
Gillam	0.28302	0.16840	0.88630	**1.00000**	00:00:55

trade-off between precision and recall with much less querying costs and comparable downloading costs. This detector also competes in terms of workload until first true positive detection with less than 6 queries and about 2 downloads on average.

Text Alignment In text alignment, given a pair of documents, the task is to identify all contiguous passages of reused text between them. Table 5 shows the overall performance of eleven plagiarism detectors that implemented text alignment. Performances are measured using precision and recall at character level as well as granularity (i.e., how often the same plagiarism case is detected) and PlagDet. The detectors are ranked by PlagDet. The best performing detector is that of Sanchez-Perez *et al.* [60], closely followed by the detectors of Oberreuter and Eiselt [41] and Palkovsii and Belov [43]. The detailed performances of each detector with regard to different kinds of obfuscation can be found in [55].

4 Conclusions

In this paper we introduced the reader to author profiling and plagiarism detection as well as to the PAN shared tasks. The difficulties of both tasks have been highlighted together with the way the participating teams have approach them. To improve the reproducibility of shared tasks, participants are asked at PAN to submit running softwares instead of their run output. To deal with the organisational overhead involved in handling software submissions, the TIRA web platform [17] helps to significantly reduce the workload for both participants and organizers, whereas the submitted softwares are kept in a running state. This year, 57 softwares have been submitted to our lab, and together with the 58 software submissions of last year, this forms the largest collection of softwares for our three tasks to date, all of which are readily available for further analysis.

In the future it would be interesting to approach author profiling in social media considering simultaneously several aspects such as gender, age and personality. With respect to plagiarism detection, recently it has been approached also in source code [12, 13], and a PAN shared task on the detection of SOurce COde (SOCO) re-use has been organised at the Forum for Information Retrieval Evaluation.[10]

Acknowledgements We would like to thank Yuri Chekhovich (Forecsys) and Mikhail Alexandrov (Russian Presidential Academy of national economy and public administration) for providing plagiarised cases in Russian for the APPD course at RuSSIR. We thank Irina Chugur (UNED) and Francisco Rangel (Autoritas Consulting) for helping with the author profiling corpus in Russian. The PAN shared tasks on author profiling and on plagiarism detection have been organised in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the EC FP 7 Marie Curie People. The research work described in the paper was carried out in the framework of the DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. Finally, we thank Hugo Jair Escalante (INAOE) for helping to improve this paper.

Bibliography

1. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.
2. Association of Teachers and Lecturers. School work plagued by plagiarism - ATL survey. Technical report, Association of Teachers and Lecturers, London, UK, 2008. Press release.
3. A. Barrón-Cedeño. *On the mono- and cross-language detection of text re-use and plagiarism.* PhD thesis, Universitat Politènica de València, 2012.
4. A. Barrón-Cedeño, P. Rosso, D. Pinto, and A. Juan. On cross-lingual plagiarism analysis using a statistical model. In *Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. PAN'08*, 2008.
5. A. Barrón-Cedeño, P. Gupta, and P. Rosso. Methods for cross-language plagiarism detection. *Knowledge-Based System*, 50:11–17, 2013.
6. A. Barrón-Cedeño, M. Vila, M. Martí, and P. Rosso. Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947, 2013.
7. D. Bogdanova, P. Rosso, and T. Solorio. Exploring high-level features for detecting cyberpedophilia. *Computer Speech & Language*, 28(1):108–120, 2014.
8. M. Braschler and D. Harman, editors. *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy, September 2010.
9. L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors. *CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/ Vol-1180/*, 2014.

[10] http://www.dsic.upv.es/grupos/nle/soco/

10. R. Comas, J. Sureda, C. Nava, and L. TITLE = Serrano.
11. R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32 (3):221–233, 1948.
12. E. Flores, A. Barrón-Cedeño, P. Rosso, and L. Moreno. Desocore: Detecting source code re-use across programming languages. In *Proc. 12th Int. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-2012*, pages 1–4, Montreal, Canada, 2012.
13. E. Flores, A. Barrón-Cedeño, L. Moreno, and P. Rosso. Uncovering source code re-use in large-scale programming environments. *Computer Applications in Engineering and Education*, Accepted, 2014. DOI: 10.1002/cae.21608.
14. P. Forner, R. Navigli, and D. Tufis, editors. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, 2013.
15. M. Franco-Salvador, P. Gupta, and P. Rosso. Cross-language plagiarism detection using a multilingual semantic network. In *Proc. of the 35th European Conference on Information Retrieval (ECIR'13)*, volume 7814 of *LNCS*, pages 710–713. Springer-Verlag, 2013.
16. M. Franco-Salvador, P. Gupta, and P. Rosso. Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In *Revised Papers PROMISE Winter School 2013*, volume 8173 of *LNCS*, pages 227–236. Springer-Verlag, 2013.
17. T. Gollub, B. Stein, and S. Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, August 2012. ISBN 978-1-4503-1472-5. doi: http://dx.doi.org/10.1145/2348283.2348501.
18. T. Gollub, M. Hagen, M. Michel, and B. Stein. From keywords to keyqueries: Content descriptors for the web. In 36th International ACM Conference on Research and Development in Information Retrieval (SIGIR 13), editors, *Gurrin, C. and Jones, G. and Kelly, D. and Kruschwitz, U. and de Rijke, M. and Sakai, T. and Sheridan, P.*, pages 981–984. ACM, 2013.
19. S. Goswami, S. Sarkar, and M. Rustagi. Stylometric analysis of bloggers' age and gender. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press, 2009.
20. G. Gressel, P. Hrudya, K. Surendran, S. Thara, Aravind A., and Prabaharan. P. Ensemble Learning Approach for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
21. C. Grozea and M. Popescu. ENCOPLOT - performance in the Second International Plagiarism Detection Challenge lab report for PAN at CLEF 2010. In Braschler and Harman [8].

22. C. Grozea, C. Gehl, and M. Popescu. ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. In Stein et al., editor, *Overview of the 1st International Competition on Plagiarism Detection*, pages 10–18, 2009.
23. R. Gunning. *The technique of clear writing.* McGraw-Hill Int. Book Co, 1952.
24. P. Gupta, A. Barrón-Cedeño, and P. Rosso. Cross-language high similarity search using a conceptual thesaurus. In *3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics. CLEF 2012*, volume 7488 of *LNCS*, pages 67–75. Springer-Verlag, 2012.
25. A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.
26. IEEE. A Plagiarism FAQ. http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html, 2008. Published: 2008; Last accessed 25/Nov/2012.
27. M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.
28. Y. Liau and L. Vrizlynn. Submission to the author profiling competition at pan-2014. In *Proc. Recent Advances in Natural Language Processing III.* http://www.webis.de/research/events/pan-14, 2014.
29. A. P. Lopez-Monroy, M. Montes-Y-Gomez, H. J. Escalante, L. Villaseñor-Pineda, and E. Villatoro-Tello. INAOE's Participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. In Forner et al. [14].
30. A. Pastor López-Monroy, M. Montes y Gómez, H. J. Escalante, and L. Villaseñor-Pineda. Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
31. S. Maharjan, P. Shrestha, and T. Solorio. A Simple Approach to Author Profiling in MapReduce—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
32. J. Marquardt, G. Fanardi, G. Vasudevan, M.F. Moens, S. Davalos, A. Teredesai, and M. De Cock. Age and Gender Identification in Social Media—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
33. B. Martin. Plagiarism: policy against cheating or policy for learning? *Nexus (Newsletter of the Australian Sociological Association)*, 16(2):15–16, 2004.
34. P. Mcnamee and J. Mayfield. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1):73–97, 2004.
35. M. Meina, K. Brodzinska, B. Celmer, M. Czokow, M. Patera, J. Pezacki, and M. Wilk. Ensemble-based Classification for Author Profiling Using Various Features—Notebook for PAN at CLEF 2013. In Forner et al. [14].
36. S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. In *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 565–569. Springer-Verlag, 2006.

37. M. Montes y Gómez, A.F. Gelbukh, A. López-López, and R.A. Baeza-Yates. Flexible comparison of conceptual graphs. In *Proc. DEXA*, pages 102–111, 2001.
38. R. Navigli and S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
39. R. M. A. Nawab, M. Stevenson, and P. Clough. University of sheffield lab report for pan at clef 2010. In Braschler and Harman [8].
40. D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "how old do you think i am?"; a study of language and age in twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
41. G. Oberreuter and A. Eiselt. Submission to the 6th international competition on plagiarism detection. http://www.webis.de/research/events/pan-14 (2014), From Innovand.io, Chile, 2014.
42. F.J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
43. Y. Palkovskii and A. Belov. Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
44. J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.
45. M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso. An evaluation framework for plagiarism detection. In *COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics*, pages 997–1005.
46. M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Proc. 30th European Conference on IR Research (ECIR)*, volume 4956 of *LNCS*, pages 522–530. Springer-Verlag, 2008.
47. M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso. Overview of the 1st international competition on plagiarism detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and E. Agirre, editors, *Proceedings of the SEPLN09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9, 2009. CEUR-WS.org (Sept 2009), http://ceur-ws.org/Vol-502.
48. M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso. Overview of the 2nd International Competition on Plagiarism Detection. In Braschler and Harman [8].
49. M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso. Overview of the 2nd international competition on plagiarism detection. In M. Braschler, D. Harman, and E. Pianta, editors, *Working Notes Papers of the CLEF 2010 Evaluation Labs (Sep 2010)*, 2010. http://www.clef-initiative.eu/publication/working-notes.

50. M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62, 2011.
51. M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein, and P. Rosso. Overview of the 3rd international competition on plagiarism detection. In V. Petras, P. Forner, and P. Clough, editors, *Working Notes Papers of the CLEF 2011 Evaluation Labs (Sep 2011)*, 2011. http://www.clef-initiative.eu/publication/working-notes.
52. M. Potthast, T. Gollub, M. Hagen, J. Grabegger, J. Kiesel, M. Michel, A. Oberlander, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso, and B. Stein. Overview of the 4th international competition on plagiarism detection. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *Working Notes Papers of the CLEF 2012 Evaluation Labs (Sep 2012)*, 2012. http://www.clef-initiative.eu/publication/working-notes.
53. M. Potthast, M. Hagen, B. Stein, J. Grabegger, M. Michel, M. Tippmann, and C. Welsch. Chatnoir: A search engine for the clueweb09 corpus. In B. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, page 1004, 2012.
54. M. Potthast, T. Gollub, M. Hagen, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein. Overview of the 5th international competition on plagiarism detection. In Forner et al. [14].
55. M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, and B. Stein. Overview of the 6th International Competition on Plagiarism Detection. In Cappellato et al. [9].
56. B. Pouliquen, R. Steinberger, and C. Ignat. Automatic linking of similar texts across languages. In *Proc. Recent Advances in Natural Language Processing III*, pages 307–316. RANLP'03, 2003.
57. A. Prakash and S. Saha. Experiments on Document Chunking and Query Formation for Plagiarism Source Retrieval-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
58. F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the Author Profiling Task at PAN 2013—Notebook for PAN at CLEF 2013. In Forner et al. [14].
59. F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkman, B. Stein, B. Verhoeven, and W. Daelemans. Overview of the 2nd Author Profiling Task at PAN 2014—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
60. M. Sanchez-Perez, G. Sidorov, and A. Gelbukh. A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
61. J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI, 2006.
62. E. Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and E. Agirre,

editors, *Proceedings of the SEPLN09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 38–46, 2009. CEUR-WS.org (Sept 2009), http://ceur-ws.org/Vol-502.

63. B. Stein, S. Meyer zu Eissen, and M. Potthast. Strategies for retrieving plagiarized documents. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *30th International ACM Conference on Research and Development in Information Retrieval (SIGIR 07)*, pages 825–826. ACM, 2007.
64. B. Stein, M. Potthast, P. Rosso, A. Barrón-Cedeño, E. Stamatatos, and M. Koppel. Fourth international workshop on uncovering plagiarism, authorship, and social software misuse. In *ACM SIGIR Forum*, volume 45, pages 45–48, 2011.
65. R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. In *Proc. 5th Int. Conf. on language resources and evaluation LREC'2006*, 2006.
66. S. Suchomel and M. Brandejs. Heterogeneous Queries for Synoptic and Phrasal Search-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
67. J. Villena-Román and J. C. González-Cristóbal. DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
68. P. Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *Int. Journal of Lexicography*, 17, 2004.
69. H. Wang, Y. Lu, and C. Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 783–792, 2010.
70. Edson R.D. Weren, V. P. Moreira, and J. P.M. de Oliveira. Exploring Information Retrieval features for Author Profiling—Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
71. K. Williams, H.H. Chen, and C. Giles. Supervised Ranking for Plagiarism Source Retrieval-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].
72. G. Yule. *The statistical study of literary vocabulary.* Camb. Univ. Press, 1944.
73. D. Zubarev and I. Sochenkov. Using Sentence Similarity Measure for Plagiarism Source Retrieval-Notebook for PAN at CLEF 2014. In Cappellato et al. [9].