Computational approaches to plagiarism detection and authorship attribution in real forensic cases

Maria Teresa Turell & Paolo Rosso [*]

Abstract. *This paper integrates work done in the fields of forensic and computational linguistics by applying computational approaches to plagiarism detection and authorship attribution in real forensic cases. It reports on findings obtained through the application of computational tools, which have been useful in plagiarism detection and analysis of real forensic cases, and computer-aided queries of annotated corpora, which have allowed forensic linguists to test the statistical significance of new morpho-syntactic markers of forensic authorship attribution such as non discrete linguistic variables (i.e. Morpshosyntactically Annotated Tag Sequences) occurring in fairly long texts.*

Keywords: Forensic linguistics, computational linguistics, plagiarism detection, authorship attribution, real forensic cases.

Introduction

One of the many areas of study of forensic linguistics has to do with written text comparison used to identify unique and idiosyncratic parameters of an individual's idiolectal style, working with the assumption that language can reveal a writer's socio-individual and socio-collective factors (age, gender, occupation, education, religion, geographical origin, ethnicity, race and language contact situation) and focusing on what the texts say and whether two or more texts have been written by the same author (authorship) or have been plagiarised from each other (plagiarism). Over the last two decades, forensic linguists have been claiming their ability to assist the court in civil and criminal proceedings and act as expert witnesses by meeting different evidentiary requirements depending on the diverse legal systems (Common and Civil Law).

[*] ForensicLab, Universitat Pompeu Fabra & Natural Language Engineering Lab, Universitat Politècnica de València

However, forensic linguists nowadays face two fundamental challenges. One has to do with the nature of written forensic texts, either long, but unique, in the sense that there may not be any other texts to compare with - so that there is only linguistic evidence within the text itself in order to establish possible plagiarism and/or come up with a linguistic profile in authorship attribution -, or very short texts (handwritten letters, annonymous electronic or type-written documents), for which the empirical evaluation of markers of disputed authorship is not easily allowed.

The second challenge is framed around the verification that the major part of the forensic written text comparison conducted these days is still quite unsystematic and unreliable. Therefore, there is a need to subject it to scrutiny in terms of methodologically-incorrect selections of the universe of non-disputed texts, ignorance of base rate knowledge information, lack of scientific design and analytical methods and, above all, in terms of the existing speculation as to the actual reliability that should be involved in evaluating whether there is more inter-writer than intra-writer variation and, furthermore, whether or not an individual's idiolectal style varies throughout his life span and beyond different genres.

Notwithstanding, during the last decade, scientific and reliable approaches to the kind of text comparison involved in plagiarism detection and authorship attribution, both in forensic and non-forensic cases, have responded to the need to rise to the challenges mentioned above. These approaches include stylometric measurements of an individual's style (Baayen *et al.*, 1996; Love, 2002; Feiguina and Hirst, 2007; Spassova and Turell, 2007; Grant, 2007; Chaski, 2001); identification of idiolectal styles (Chaski, 2001; Grant and Baker, 2001); stylistic methods (McMenamin, 2001), and vocabulary analytical techniques (Coulthard, 2004; Turell, 2004), with consideration of core lexical elements, hapax legomena, hapax dislegomena, lexical density and lexical richness and the use of corpus of reference in order to establish the low/high frequency of words in disputed text sets, by taking into account the concepts of markedness and saliency.

Several are the investigations carried out in order to deal with the above-mentioned difficulties and problems from a computational linguistics perspective. In Shivakumar and Garcia-Molina (1995) a copy detection approach based on word frequency analysis was introduced. In Kang *et al.* (2006) an approach based on word comparison at sentence level which takes into account vocabulary expansion with Wordnet[1] was described. A few methods attempt to solve plagiarism detection on the basis of word n-grams comparisons (Lyon *et al.*, 2004; Muhr *et al.*, 2010) or also character n-grams (Schleimer *et al.*, 2003; Grozea *et al.*, 2009). Recently, researchers have also approached the issue of cross-language plagiarism (Potthast *et al.*, 2011; Barrón-Cedeño *et al.*, 2010; Gupta *et al.*, 2012; Franco-Salvador *et al.*, 2013). When there may not be any suspicious text to compare the suspicious document with, the linguistic evidence may have to be provided on the basis of stylistic changes found in the document itself (intrinsic plagiarism) (Stein and Meyer zu Eissen, 2007). Another method for intrinsic plagiarism detection is the one described in Stamatatos (2009), where character n-gram profiles have been used. Computational linguists have also considered the somewhat related issue of authorship attribution, where linguistic profiles need to be investigated in order to try to determine who the real author of a text is (Stamatatos *et al.*, 2000; Koppel *et al.*, 2009).

This article argues that forensic plagiarism detection and authorship attribution can benefit from the complementary interplay between approaches used in forensic linguis-

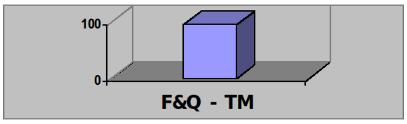


Figure 1. Overlapping vocabulary (F&Q-TM) Activities (Turell, 2008: 288)

tics (textual qualitative analysis, observation of semantic and pragmatic markers that cannot be analysed by automatic procedures, the use of reference corpora to set up the rarity or expectancy of the writers' idiolectal choices) and those used in computational linguistics, automatic or semiautomatic natural language processing techniques that would allow researchers to establish the statistical significance of results, something which is becoming more and more necessary when acting as linguistic expert witnesses in court. The studies presented in this article report on findings from the application of both computational tools and computer-aided queries of annotated corpora. They have been useful in plagiarism detection and analysis of real forensic cases, allowing forensic linguists to test the statistical significance of new morpho-syntactic markers of forensic authorship attribution such as non discrete linguistic variables (i.e. Morpshosyntactically Annotated Tag Sequences) occurring in fairly long texts.

In-tandem forensic and computational approaches to plagiarism detection

Forensic linguistics makes a distinction between copying of ideas and linguistic plagiarism. Copying of ideas can exist without linguistic plagiarism but if the latter is detected, the former occurs as well by the nature and definition of the linguistic sign itself. This discipline has devised a methodology which includes several qualitative and quantitative tools and is used to establish a) the nature and degree of plagiarism, b) plagiarism directionality, and c) the threshold level of textual similarity between texts above which this similarity becomes suspicious. For the purposes of this article, we will consider point c) in particular and explain what tools are used to analyse this textual similarity.

CopyCatch[2], one of the many existing concordance tools used to detect linguistic plagiarism, allows researchers to calculate the threshold level of textual similarity which becomes suspicious in a forensic case. This program incorporates several measurements such as threshold of overlapping vocabulary, hapax legomena, hapax dislegomena, unique and exclusive vocabulary and shared-once phrases. It has a visual output, with plagiarised sections marked in red, which is very useful when presenting evidence in court. In order to establish this threshold level of textual similarity forensic linguists can count on empirical evidence which suggests that "up to 35% similarity is normal and up to 50% is not unusual, although the further above 50% the more likely it is to indicate that the texts under consideration have not been produced independently and that there exists a borrowing relationship between the texts under consideration". Empirical research has also proved that this threshold level should be increased up to 70%, in the case of plagiarism between translations (Turell, 2004).

Out of the eight (8) Spanish plagiarism detection cases, in which our forensic linguistics laboratory has been involved in the last 7 years, three (3) present verbatim plagiarism and the other five (5) reflect partial plagiarism, with varying degrees of paraphrase and overlapping vocabulary. The domains of occurrence of these cases are education (text

Table 1. Decontextualization of F&Q (Bruño) in TM (Temario Magister) (Turell, 2008: 286)

Física y Química Bruño (2002) (page 194)	**Temario Magister (2005) (Unit 11, pages 8 & 9)**
ACTIVITY 2	**ACTIVITY 1**
2. Completa en tu cuaderno la siguiente Tabla 11.2 consultando la Tabla Periódica.	1. Completa la siguiente Tabla:

ENGLISH TRANSLATION

Física y Química Bruño (2002) (page 194)	**Temario Magister (2005) (Unit 11, pages 8 & 9)**
ACTIVITY 2	**ACTIVITY 1**
2. Complete in your exercise book the following Table 11.2 consulting the Periodic Table.	1. Complete the following Table:

books), tourism (guides and brochures), scientific research, music (lyrics) and literature (novels). For example, Figure 1 shows the threshold level of overlapping vocabulary (96%) found when comparing the sections on Activities in the two text books under analysis: the non-disputed text Física y Química Bruño (F&Q, 2002) and the disputed text Temario Magister (TM, 2005) (Bruño vs. Magister),a percentage which indicates that Activities in F&Q have been reproduced almost verbatim in TM.

One of the other equipment facilities supplied by CopyCatch is that the program takes you to the Sentences Page automatically. This Sentences Page, which presents verbatim, or almost verbatim, phrases/sentences, facilitates the identification of the uncoherent/uncohesive segments and the plagiarist's strategies within the whole text, which may lead to a) meaningless sequences due to the 'cut & paste' technique used, b) inconsistency in referential style, c) decontextualisation and d) inversion in the grading of structural elements, among others.

Table 1 illustrates one example of decontextualisation produced by the fact that one part of the directions given in Activity 2 (page 194) in the non-disputed text (F&Q, 2002), namely, "consultando la Tabla Periódica", has been deleted in Activity 2 (Unit 11, page 8) in the disputed text (TM, 2005).

Pl@giarism[3], developed by the University of Maastricht, is another system used in plagiarism detection. The system returns the percentage of textual similarity between two documents (A and B), the percentage of the number of matches with document A versus document B, the percentage of the number of matches with document B versus document A and the total amount of matches between documents A and B. This system performs the comparison on the basis of word trigrams. Like CopyCatch, Pl@giarism has a visual output, with plagiarized sections marked in red. WCopyFind[4], developed by the University of Virginia, is another tool made available for plagiarism detection (Vallés Balaguer, 2009). The system allows researchers to introduce various parameters such as the size of word n-grams, the minimum number of matching words to be reported as possible plagiarism, the maximum number of non matches between perfectly matching portions of a sentence, etc. Apart from highlighting the substitution of words

Table 2. Comparison of results: CopyCatch v. WCopyFind v. Pl@giarism

	CopyCatch	WCopyFind	Pl@giarism
Case: Bruño v. Magister			
Page 5	27 %	19 %	21 %
Page 32	79 %	92 %	92 %
Pages 33-37	95 %	96 %	96 %
Pages 40-46	94 %	86 %	83 %
Case: XXX (for anonymity) v. Signes			
Activities	96 %	87 %	86 %
Cuestionario	98 %	87 %	78 %
Técnico	86 %	83 %	78 %

by synonyms, one added value of WCopyFind is the provision of a word map (a generalized thesaurus). This system tells researchers the percentage of the number of matches with document A versus document B, the percentage of the number of matches with document B versus document A, and like CopyCatch and Pl@giarism, WCopyFind has a visual output, with plagiarized sections marked in red.

Table 2 shows a comparison of CopyCatch, WcopyFind and Pl@giarism results related to the detection of plagiarized fragments found in the corpus sets of two real forensic cases (Bruño vs. Magister and XXX v. Signes. In WCopyFind we used trigrams as n-gram size, and five as the maximum number of non matches. Results illustrate that the three tools have been able to detect plagiarism in these real cases. However, in most cases, the tool CopyCatch returns the highest percentage of similarity between the texts compared. This is because this tool runs on unigrams, once-shared words and unique vocabulary. WCopyFind returns a higher percentage than Pl@giarism, the main reason for this being that WcopyFind considers the possible words inserted between perfectly matching sentence fragments, whereas Pl@giarism does not.

For real cases Bruño vs. Magister and XXX v. Signes, some linguistic evidence was given on the basis of the comparison with other texts. However, as mentioned above, one of the challenges forensic linguists have to face is that very often written forensic texts are unique in the sense that there may not be any other texts against which to compare them. Therefore, in order to be able to establish possible plagiarism the linguistic evidence has to be found within the text itself (intrinsic plagiarism). YYY[5] is a new tool the aim of which is to help forensic linguists to come up with a linguistic profile on the basis of a stylistic analysis of the text in order to determine whether or not there are fragments of different writing styles. This tool divides the text into fragments and for each of these fragments, it calculates various vocabulary richness measurements (function K proposed by Yule (Yule, 1944), function R proposed by Honore (Honore, 1979)) and text complexity (Flesch-Kincaid Readability Test (Flesch, 1948), Gunning Fog Index (Gunning, 1952)) as suggested in Meyer zu Eissen *et al.* (2007). The aim behind Stylysis is to identify text fragments with different writing styles, which could indicate that it is a plagiarized fragment or that it has been written by a different author. Thus, this tool could also help in linguistic profiling to attribute authorship of a text.

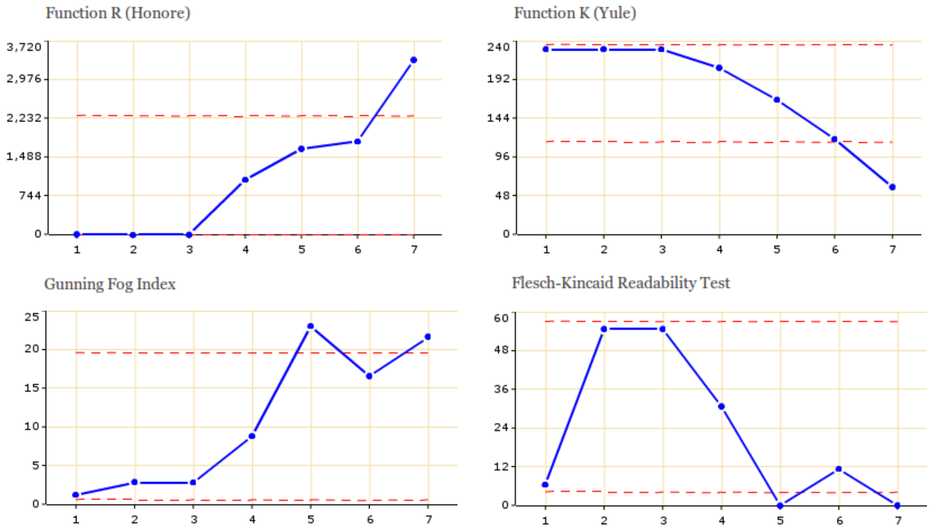


Figure 2. Results obtained with the Stylysis tool in case XXX v. Signes (5)

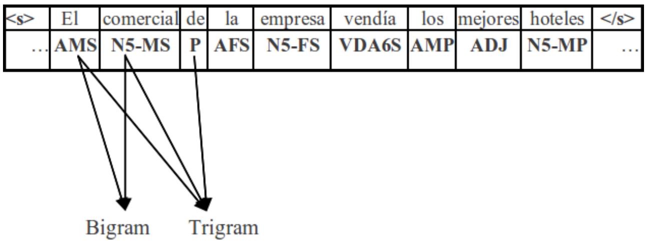
Figure 2 shows the results for page 5 of the corpus set of case XXX v. Signes. R and K functions show that fragment 7 exceeds the standard deviation (dashed line). This could indicate a possible change in writing style. The Flesch-Kincaid Readability Test and the Gunning Fog Index show that fragments 5 and 7 exceed the standard deviation. On the basis of these measurements, fragment 7 is the only suspicious fragment. And indeed fragment 7 proved to be a case of plagiarism. This example demonstrates the usefulness of Stylysis in providing language evidence for intrinsic plagiarism detection.

The use of computer-aided queries of annotated corpora in forensic authorship attribution

One computational approach used in the kind of forensic text comparison leading to more reliable authorship attribution outcomes is the study of the syntactic characterization of a writer's idiolectal style through computer-aided queries of annotated corpora that can help to establish the statistical significance of sequences of linguistic categories, namely, Morpho-syntactic Annotated Tag Sequences (MATS), as proposed in Spassova and Turell (2007). This approach is not new in non-forensic contexts (Baayen *et al.*, 1996), where these sequences are frequenly referred to as n-grams (and depending on the number of categories combined, the terms used are bigrams, trigrams, etc.), but the ForensicLab has been one of the first to apply this method to real forensic cases.

This method is structured around the following activities:

1. A pre-processing phase, in which texts are segmented into their basic components: title, paragraphs, sentences, and paragraph beginnings and ends are marked ($< /s >< /p >$).
2. A morpho-syntactic tagging phase, during which the text is converted into a flow of token types and tags.
3. A disambiguation stage, through which texts are disambiguated and errors are corrected.
4. A tag extraction phase - making use of LEGOLAS 2.0 - during which the information obtained refers to the number of MATS types and tokens and on the MATS frequency values to be used in the subsequent statistical analysis.



<s>	El	comercial	de	la	empresa	vendía	los	mejores	hoteles	</s>
…	AMS	N5-MS	P	AFS	N5-FS	VDA6S	AMP	ADJ	N5-MP	…

Figure 3. Morpho-syntactically Annotated Tag Sequences (MATS)

5. Once the tags have been extracted, a last stage involves the application of Discriminant Function Analysis (DFA), in order to classify the different text sets, and the projection of results onto graphs.

During the pre-processing and processing phases several processing and disambiguation tools from the IULA's technical corpus[6] were used (Morel *et al.*, 1998). For example, once tagged, the sentence: "el comercial de la empresa vendía los mejores hoteles" ('the firm's salesperson was selling the best hotels') is projected and represented in the way Figure 3 illustrates. Two examples of MATS are marked, namely, "el comercial" (AMS N5-MS), which is a bigram and "el comercial de", (AMS N5-MS P), which is a trigram, and where A stands for article, M for masculine, S for singular, N5-MS for singular masculine common noun, and P for preposition.

Part of the linguistic evidence used to report on six (6), out of nine (9), Spanish real forensic authorship attribution cases considered was drawn by applying this methodological protocol. Outcoming results from all these cases have shown that the discriminatory potential of MATS is higher with long text samples and with a big number of control reference texts (Grant, 2007) and that bigrams and trigrams are more discriminatory than longer sequences of linguistic categories. To test the working hypotheses which are at play in forensic written text comparison - that is, a) that everyone has an 'idiolectal style', as relating to the linguistic system shared by lots of people, but used in a distinctive and unique way by particular individuals, who have different options at their reach in their linguistic repertoire and make selections from these options (Halliday, 1989), and b) that a disputed text (or several disputed texts) can be attributed to the same author who wrote a set of non-disputed texts - sets of anonymous texts from other real forensic cases are used, thus optimizing the discriminatory potential of MATS.

Figure 4 shows the projection for bigrams of the DFA applied to the three text sets under analysis in the forensic case XXX v. SEHRS: the disputed e-mails (+), the non-disputed faxes (△) and the anonymous emails from another case (◯, indicated as docs in Figure 4). In this figure, it can be seen that although there is a certain distance between the disputed emails and the non-disputed faxes, the distance of all the emails from another case and the text sets relevant to the case under analysis is even bigger, which indicates more statistically significant differentiation in the idiolectal use of MATS in the emails from another case than the one found in the comparison of the NDTfax and DT@ text sets.

The classification method of DFA classified with success 100% of the texts by authors within their own group while the cross-validation method confirmed that the analysis was 83.4% correct. Two of these emails belong to the non-disputed faxes set, whereas the

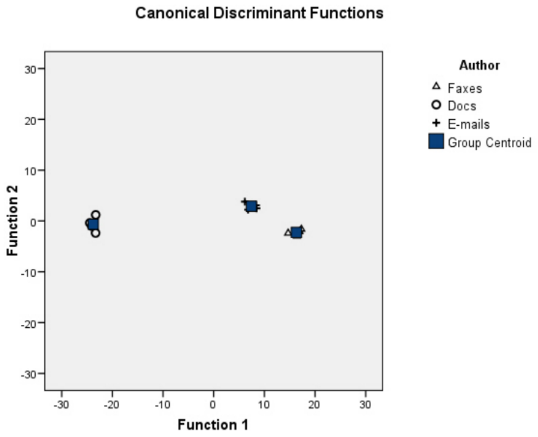


Figure 4. Discriminant Function Analysis (NDTfax and DT@) - Bigrams (Turell, 2010)

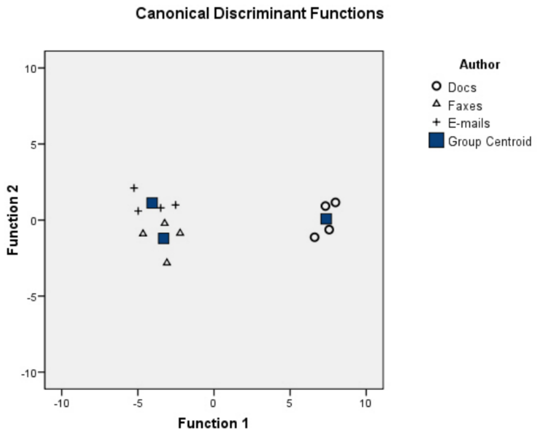


Figure 5. Discriminant Function Analysis (NDTfax and DT@) - Trigrams (Turell, 2010)

other two disputed emails are classified within its original group, and the anonymous emails from another case are all classified within their original group. This outcome seems to confirm that the probability that the author of the disputed emails could be the author of the non-disputed faxes is quite high.

Figure 5 shows the projection of the results for trigrams. This figure illustrates that the centroids of the disputed emails (+) and the non-disputed faxes (△) are placed in the same area of the graph, while the centroid of the emails from another case (◯) is located in the opposite side of the graph.

However, on this occasion the DFA classification method shows that only 75% of the texts are classified with success; only three of the disputed emails are classified as if produced by the author of the non-disputed faxes; besides, cross-validation confirmed that the analysis was only 63% correct, since two of the disputed emails are attributed to the group of non-disputed faxes and two of the non-disputed faxes are attributed to the group of disputed emails.

These statistical results using DFA reveal that, in spite of the reduced corpus size and the short text length (although in effect the total N for MATS is not that small - 1,589 tokens for bigrams and 660 tokens for trigrams), these structures seem to exhibit a quite high discriminatory potential and also that bigrams turn out to be more discriminatory than trigrams, as other forensic cases have shown. This would allow us to conclude that sequences of grammatical categories observed in a writer's 'idiolectal style' can be used quite reliably as valid markers of authorship.

Conclusions

In this article we have attempted to show that, for both computational and forensic linguistics, joint work in the areas of forensic plagiarism detection and authorship attribution can be very fruitful. Present-day comparative methods used in forensic plagiarism detection and authorship attribution exhibit limitations; so there is a need to count on intra-evidential complementary evidence which is tested with computational (automatic or semi-automatic) natural language processing techniques. Computational linguistics, on the other hand, needs to be able to use linguistic data from real forensic cases - and not just synthetic data, automatically generated or manually generated through Amazon mechanical Turk (Potthast *et al.*, 2010) - in order to establish the actual performance features of the existing systems of automatic plagiarism detection (Barrón-Cedeño *et al.*, 2013). It is precisely because of the nature and length of forensic texts (usually quite short) and corpora (small-sized), which can be a drawback when trying to establish the statistical significance of results, that computational linguistics must come into play so that forensic linguists are able to refine their comparative methods and techniques. In this article we have reported on the comparative evaluation of three plagiarism detection tools (CopyCatch, extensively used in forensic plagiarism detection, WCopyFind and Pl@giarsm) that are available to forensic linguists, while we are aware that there are other automatic plagiarism detection systems that define the state of the art and are part of the know-how of the PAN competition on Plagiarism Detection (Potthast *et al.*, 2012).

One important empirical question that can be raised, but not answered in this article, is what kind of evaluation results would be drawn when automatic detection tools can be applied to real forensic data, once these systems are commercialized or become public domain tools. This is only one first enriching step towards the establishment of stronger collaboration links between forensic and computational linguists. However, it has not been possible to compare the forensic protocols and automatic authorship attribution techniques used in forensic linguistics with other existing automatic approaches to written authorship such as those devised by computational linguistics, which will be discussed in the context of the PAN competition on Authorship Identication[7].

Acknowledgments

"I end on a very sad note. One of the authors, Maria Teresa Turell, passed away on April 24, 2013, just before this volume went to press. Maite, as everyone knew her, was a Professor of English at Universitat Pompeu Fabra in Barcelona, where she directed the Forensic Linguistics Laboratory. She was an important figure in the field, devoted to bringing quantitative rigor to stylistic insight, one of the themes of the workshop. More importantly to those of us who knew Maite, her intellectual toughness was matched with a loving and generous character that will remain with us for a long, long time to come." Authorship Attribution Workshop, Journal of Law and Policy 21(2). Brooklyn, N.Y. Preface by Lawrence M. Solan, Brooklyn Law School

Cases cited

Case: Bruño vs. Magister.
Case: XXX v. Signes.
Case: XXX vs. SEHRS.

Notes

[1] http://wordnet.princeton.edu

[2] CopyCatchGold, CFL Development, http://cflsoftware.com/

[3] http://www.plagiarism.tk

[4] http://www.plagiarism.phys.virginia.edu/Wsoftware.html

[5] http://memex2.dsic.upv.es:8080/StylisticAnalysis/es/index.jsp

[6] http://brangaene.upf.es/plncorpus/index2.htm, http://brangaene.upf.es/plncorpus/faq.html

[7] http://pan.webis.de/

References

Baayen, R. H., van Halteren, H. and Tweedie, F. J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121–131.

Barrón-Cedeño, A., Rosso, P., Agirre, E. and Labaka, G. (2010). Plagiarism detection across distant language pairs. In *Proc. of the 23rd International Conference on Computational Linguistics, COLING-2010*, 37–45, Beijing, China.

Barrón-Cedeño, A., Vila, M., Martí, A. and Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4).

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 8(1), 1–65.

Coulthard, R. M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4), 431–447.

Feiguina, O. and Hirst, G. (2007). Authorship attribution for small texts: Literary and forensic experiments. In *International Workshop on Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection, PAN 2007*: 30th Annual International ACM SIGIR Conference.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

Franco-Salvador, M., Gupta, P. and Rosso, P. (2013). Cross-language plagiarism detection using multilingual semantic network. In *Proceedinfs of 35th European Conference on Information Retrieval, ECIR-2013*, Moscow, Russia.

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law*, 14(1), 1–25.

Grant, T. and Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8(1), 66–79.

Grozea, C., Gehl, C. and Popescu, M. (2009). Encoplot: Pairwise sequences matching in linear applied to plagiarism detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 10–18, San Sebastian, Spain.

Gunning, R. (1952). *The technique of clear writing.* McGraw-Hill Int. Book Co.

Gupta, P., Barrón-Cedeño, A. and Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In *Proceedings of 3rd International Conference of CLEF on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics, CLEF 2012*, 67–75, Rome, Italy.

Halliday, M. (1989). *Language, context and text. Aspects of language in a social semiotic perspective.* Oxford: Oxford University Press.

Honore, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2), 172–177.

Kang, N., Gelbukh, A. and Han, S. Y. (2006). Ppchecker: Plagiarism pattern checker in document copy detection. *Lecture notes in computer science.*, 4188, 661–668.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Love, H. (2002). *Attributing authorship: An introduction.* Cambridge: Cambridge University Press.

Lyon, C., Barrett, R. and Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. In *Plagiarism: Prevention, Practice and Policies Conference*, Newcastle, UK.

McMenamin, G. (2001). Style markers in authorship studies. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 8(2), 93–97.

Meyer zu Eissen, S., Stein, B. and Kulig, M. (2007). Plagiarism detection without reference collections. In R. Decker and H. Lenz, Eds., *Advances in Data Analysis*, 359–366.

Morel, J., Torner, S., Vivaldi, J. and Cabré, M. T. (1998). *El corpus de l'IULA: Etiquetaris.* Serie Informes, 18 (2nd Edition). Barcelona: Publicacions de l'IULA.

Muhr, M., Kern, R., M., Z. and Granitzer, M. (2010). External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In *Notebook Papers of CLEF 2010 LABs and Workshops*, University of Padova, Italy.

Potthast, M., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING-2010*, 997–1005, Beijing, China.

Potthast, M., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2011). Cross-language plagiarism detection. *Languages Resources and Evaluation. Special Issue on Plagiarism and Authorship Analysis*, 45(1).

Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P. and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In *Notebook Papers of CLEF 2012 LABs and Workshops, CLEF-2012*, Rome, Italy.

Schleimer, S., Wilkerson, D. S. and Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY.

Shivakumar, N. and Garcia-Molina, H. (1995). Scam: A copy detection mechanism for digital documents. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries.*

Spassova, M. S. and Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S.

Spassova and J. Cicres, Eds., *Proceedings of the Second European IAFL Conference on Forensic Linguistics/Language and the Law.*, 229–237, Barcelona.

Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, 38–46.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4), 461–485.

Stein, B. and Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis with meta-learning. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, Amsterdam, Netherlands.

Turell, M. T. (2004). Textual kidnapping revisited: the case of plagiarism in literary translation. *The International Journal of Speech, Language and the Law*, 11(1), 1–26.

Turell, M. T. (2008). Plagiarism. In J. Gibbons and M. T. Turell, Eds., *Dimensions of Forensic Linguistics*, 265–299, Amsterdam/Philadelphia: John Benjamins.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17(2), 211–250.

Vallés Balaguer, E. (2009). Putting ourselves in sme's shoes: Automatic detection of plagiarism by the wcopyfind tool. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 34–35, San Sebastian, Spain.

Yule, G. (1944). *The statistical study of literary vocabulary.* Cambridge University Press.