Cross-Linguistic Stylometric Features: A Preliminary Investigation

Patrick Juola[1], George K. Mikros[2]

[1]Duquesne University + Pittsburgh – USA

[2]National and Kapodistrian University of Athens + Athens – Greece

Abstract

Stylometric analysis - the study of the writing style of the author of a document, either to determine his/her identity or personal characteristics - is an important problem in text analysis and information retrieval, with many important real-world applications. It is generally limited by a need for reference documents that are representative of the unknown documents to be analyzed. This paper addresses the issue of analysis with highly *unrepresentative* documents, and specially the question of whether elements of writing style can be shown to vary systematically with the individual irrespective of the language of writing.

We identify fourteen Twitter users who post bilingually in both Spanish and English. An analysis of several standard linguistic and Twitter-specific extra-linguistic variables show both that there is a substantial amount of individual variation along these variables, but (more importantly), that the variations correlate very strongly across languages. In other words, an individual who scores highly along one axis in English is also very likely to score highly on that axis in Spanish. These findings strongly suggest that cross-linguistic individual authorship features can be developed that, in turn, will enable accurate stylistic analysis across language barriers.

Key words : Authorship attribution, stylometry, cross-linguistic, social media.

1. Introduction

Sometimes, one has a document and needs to know not what it's about, but who wrote it -- for example, a teacher looking at a possibly plagiarized term paper, or a policeman looking at a ransom note. Authorship attribution, at JADT and elsewhere, has become a well-studied field with a well-understood stylometric methodology. As detailed below, many studies have shown the existence of systematic persistent patterns in the writing style of an individual (what Van Halteren (2007) has called the "stylome"). To address this, one gathers a set of known documents representative of and comparable to the questioned document(s), extracts a suitable feature set from the known documents and uses classification techniques to determine the author of the unknown one(s). A well-known limitation of this method is the need for comparable documents, which are often difficult to find in realistic situations.

Unfortunately, there are often practical reasons why representative documents are not available for the task at hand; few people, for example, have written an extensive library of genuine suicide notes (Chaski, 2005). This paper proposes some techniques that can be used to validate authorship based on unrepresentative training texts, and specifically in instances where the training documents and testing documents (known and questioned documents, in forensic terminology) are in entirely separate languages. By analyzing a corpus of such texts, gathered from social media, we will show that some of the features used in ordinary, single-language cases, can and do persist across languages as well.

2. Background

2.1. Subsections

A recent high-profile example of authorship attribution (Brooks, 2013; Brooks & Flyn, 2013) is that of the author of *A Cuckoo's Calling*, by Robert Galbraith. Formal analyses of writing style, performed at the behest of the *Sunday Times*, later identified J.K. Rowling, author of the Harry Potter books, as the actual author (Juola, 2013a). Literature scholars have been interested in questions of authorship for centuries, but identifying the author of a document can be of interest to other parties as well.

A non-obvious but key application is to the legal system. For example, a famous dispute over the ownership of a significant part of Facebook (Ceglia vs. Zuckerberg and Facebook) depended in part upon a set of disputed writings. McMenamin (2011) submitted a report in this case that showed that the writing style of a set of undisputed email (that Zuckerberg acknowledged having written) differed in a number of important ways from the disputed writings, and concluded that "[i]t is probable that Mr. Zuckerberg is not the author of the QUESTIONED writings." (Capitalization in original.) Other court cases related to actual authorship disputes include (Chaski, 2005; Coulthard, 2013; Grant, 2013; Juola, 2013b). From fraud to murder, the legal applications of stylometry are significant.

Another common application is journalism. As with the Rowling case (Brooks & Flyn), many questions arise from a matter of public interest, driven by journalists. Another recent example is Newsweek's analysis of the Bitcoin design documents, attributed by Newsweek to a retired engineer named Dorian Nakamoto. Stylometric analysis of these documents (Herper) against an appropriate set of known documents showed "that Dorian Nakamoto was not found to be a plausible candidate author, and in fact, one of the distractor authors (Neal J. King) was found to be a better match to Satoshi Nakamoto than any other distractor or than Dorian." (Juola, 2014). An open problem is that of Fuat Avni (Kocagul, 2014), an anonymous critic on social media of the current Turkish prime minister. "He reveals interesting anecdotes as a close associate of [the current Turkish] government" and has been suggested to be "the reason Erdogan banned Twitter." However, no one knows who he (or she) is, or even whether s/he is a single author or many.

2.2. Theory of Stylometry

So how does this work? The basic theory of traditional stylistics is fairly simple. As McMenamin describes it,

> At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer's "choice" of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer's own unique set of habitual linguistic choices (McMenamin, 2011).

Recent scholarship has established that higher performance can generally be obtained by using low-level and linguistically unsophisticated feature sets such as word choice (Binongo, 2003; Burrows, 1989; Hoover, 2004) or even character clusters (Juola et al., 2013; Mikros & Perifanos, 2013; Stamatatos, 2013). McMenamin's report, for example analyzed eleven different and distinct aspects of the writing in both the known (undisputed) email and the

disputed email. One feature hinged on the spelling of the word cannot, and in particular whether it was written as one word *(cannot)* or as two *(can not)*. Another feature was the use of the single word "sorry" as a sentence opener (as opposed, for example, to "I'm sorry"). Coulthard similarly discussed (among other features) the use of the phrase "disgruntled employees", while Grant's features included variant spellings such as "wiv" for "with" and "wud" for "would". Perhaps obviously, neither "wiv" nor "with" are likely to appear in a disputed document written in Spanish, and hence Grant's method will not easily transfer.

Juola's analysis of the Rowling case (Brooks & Flyn, 2013; Juola, 2013a), however, included word length (de Morgan, 1851/1882) as one of the features of analysis. While Spanish does not have the word "with", it has words in general. Is a tendency to choose long words (or short words) something that a single person would show, irrespective of the language of writing?

2.3. Twitter

As discussed below, our source for the data analyzed in this paper was Twitter. Twitter (www.twitter.com) is a social media platform focused on so-called "microblogging". Users broadcast ("blog", itself an abbreviation for "web log") short messages (of up to 140 characters), called "tweets". The Twitter platform supports a number of multimedia extensions, such as the ability to publish photographs, videos, and/or web links as well. In part due to the length restrictions, the Twitter user pool has developed some specific conventions (detailed below) that users may choose to participate in or not. More importantly for our study, Twitter is relatively language-agnostic and therefore has participants from all over the world using a variety of languages. In particular, it includes participants who use multiple languages over the course of their posting history.

Authorship attribution methods have already been applied to tweets, since Twitter is an extremely popular service and cybercrime frequently uses it for illegal activities. Examples of such studies include (Layton, Watters, & Dazeley, 2010; Mikros & Perifanos, 2013; Sousa-Silva et al., 2011). Results have shown that reliable authorship attribution results can be obtained using as little as 100 tweets per author.

Using Twitter, we focus here on two specific types of features that are not tied to any particular language (unlike, say, the appearance of a specific character or word). The first is simply "vocabulary richness", as measured by a wide variety of methods. The second is the frequency of participation in various Twitter-specific social conventions.

3. Materials & Methods

3.1. Materials: A Bilingual Twitter Corpus

We first identified (by manual inspection) a set of 16 user names that could be confirmed to have published tweets in both English and Spanish. Once our user list had been collected, we scraped the Twitter history of each user to collect tweets from each one. Each tweet, in turn, was automatically analyzed by the language detection web service provided by detectlanguage.com, to determine the language of the tweet as well as a confidence measure assessing the "reliability" of the analysis. Of the 16 users identified, one user had deleted his/her account in the interim and hence had no tweets to harvest. Another had only one Spanish tweet, and thus was dropped from further study. We restricted our attention to those

tweets that had been identified as "reliable" Spanish ("es") or English ("en") texts. Table 1 shows the distribution of tweets by user and language (specific user names have been redacted).

User	English (reliable)	Spanish (reliable)
S01	51	263
S02	49	61
S03	313	25
S04	116	218
S05	18	38
S06	280	146
S07	140	94
S08	167	654
S09	62	60
S10	47	103
S11	468	664
S12	157	127
S13	35	161
S14	20	38

Table 1: Distribution of tweets by user and language

3.2. Methods

One of the most obvious features in authorship attribution [indeed, Juola (2008) traces this idea to de Morgan (1851/1882)] is that of word length, and specifically the average (mean) length of words used in a set of writings. It is superficially plausible that people with a tendency to use long words (and hence a higher than average mean word length) would retain this tendency irrespective of the language in which they are writing. Accordingly, we tokenized all tweets and determined per-user length averages for both Spanish and English. A similar process yielded the average number of words per tweet, a measure of the length and complexity of each individual message in the stream.

Word length is often viewed as a proxy for vocabulary richness and complexity. As additional measures of vocabulary richness (another well-studied stylometric variable), we used the traditional type/token ratio (TTR), a measure of the number of times words are repeated in the text. We also noted two other traditional measures of vocabulary richness: percentage of hapax legomena (words that appear exactly once in the corpus) and Yule's K, a measurement of the likelihood that two randomly-chosen tokens are the same type.

Of course, vocabulary richness can be measured in other ways. The QUITA software package (Kubát, Matlach, & Čech, 2014) provides several additional quantitative assessments of text that are less commonly used and less-well studied. We measured, in addition, the entropy (H), R1, repeat rate (RR), RRmc, curve length (L), R index, adjusted modulus, and Gini coefficient (G) (see Kubát et al. (2014) for explanations omitted here for brevity).

In addition to these simple linguistic features, we focused on three extra-linguistic features (hashtags, mentions, and hyperlinks) that are specific to the nature of Twitter discourse. Hashtags are individual words (or unspaced phrases) prefixed with a hash (#) character that are used to identify and label messages related to a specific topic; mentions are individual

words (typically the names of other Twitter users) prefixed with a commercial at character (@) used to identify specific discourse participants, such as a person responded to or a person whom you hope will notice the message. Similarly, another common extra-linguistic practice is the inclusion of hypertext Web links in Tweets as a method of commenting on or disseminating messages that do not fit comfortably within the strict limits of tweet length. These links can often be identified as they begin with "http" and continue with a reference to a specific web page. (N.b. that not all links use the http notation.) Again, these links were identified based on the initial four characters, and, for every user, we determined the percentage of words (all links are a single word) that were such links.

3.3. Data Treatment

In an effort to control for the known sample size effects of vocabulary richness measures (especially as the available corpora varied in size by more than an order of magnitude), we tried two experiments on these measures. The first involved no corpus preprocessing. The second was simply to truncate all tweet collections to the size of the smallest one, thus studying only the first two hundred words tweeted by each individual. We then proceeded to measure the mean vocabulary richness for all subsamples corresponding to a specific author-language pair and applied simple correlation statistics (Pearson's r). As our hypothesis is that similar behaviors hold across languages, one-tailed tests are appropriate for this study, focusing on the positive tail only. All analyses were done on the fourteen (n = 14) subjects described in subsection 3.1.

Taking seriously the idea that these different richness indices measure different things, we performed a hierarchical cluster analysis of the individual indices. This shows both which indices measure the similar things and provides another check of the cross-linguistic robustness of any individual measure.

4. Results

The results of the correlation study are attached as table 2. The cluster analysis results are attached as figure 1. Note that each richness index appears in figure 1 twice; the "_SP" suffix indicates that it is the measure on the Spanish half of the corpus. (For example, RR is the repeat rate of English documents, while RR_SP is the repeat rate of Spanish ones.)

Of the features studied, almost all achieved significance on at least some conditions, and three (L, Adjusted Modulus, and G) were significant under all test conditions. Even the feature that did not reach formal significance (words per tweet) nevertheless showed evidence of a trend-level ($p < 0.10$) relationship between linguistic behavior in English and in Spanish. There was no analysis that showed no evidence supporting the existence of cross-linguistic writing style under any condition.

Examination of figure 1 supports this as well. We note first that the correlation between the various measures is extremely high, and that they generally appear, therefore to be measuring very similar things. The exceptions are a small cluster of three measures: curve length (L), Yule's K, and Redundancy, which are notably outliers. However, they are notably outliers in both English and Spanish, suggesting that whatever they specifically measure is itself stable across languages. In particular, the measurement L is stable enough across languages that the measurement of vocabulary richness via L in English is more similar to the measurement of vocabulary richness via L_SP in an entirely different language (for the same author) than it is for the measurement of vocabulary richness via any other method on, literally, the exact same

data. Whatever L specifically measures is more closely tied to the author than to the language of authorship.

	Untruncated		200 words (1st)	
Index	*r*	*p-value*	*r*	*p-value*
TTR	0.2690	p>0.1(n.s.)	0.7431489	0.002321(**)
Hapax %	0.2750575	p>0.1(n.s.)	0.7142336	0.002054 (**)
Entropy	0.3307389	p>0.1(n.s.)	0.7059129	0.002392 (**)
Lambda	0.3925311	0.08253 (n.s.)	0.6937589	0.002961(**)
Redundancy	0.3013017	p>0.1 (n.s)	0.6924124	0.00303 (**)
Popescu's R1	0.7174944	0.001933 (**)	0.3504229	p>0.1 (n.s.)
Yule's K	0.4209762	0.06694 (n.s.)	0.5906128	0.01308 (*)
RR	-0.02707195	p>0.1 (n.s.)	0.5633507	0.01796 (*)
RRmc	0.6671255	0.004576 (**)	0.2678785	p>0.1 (n.s.)
L	0.6394239	0.006903 (**)	0.6584157	0.00523 (**)
Adj. Mod.	0.5986507	0.01185 (*)	0.485988	0.03904 (*)
G	0.5021349	0.03365 (*)	0.6544689	0.005549 (**)
Curve Length R	0.5021349	0.05499 (n.s.)	0.4983632	0.03486 (*)
Characters/word	0.5902	0.0131 (*)		
Words/tweet	0.4243	0.0650 (n.s.)		
Use of #hashtags	0.8939	0.0001 (**)		
Use of @mentions	0.5826	0.0144 (*)		
Use of http: hyperlinks	0.7965	0.0003 (**)		

p < listed value unless noted
(n.s.) indicates a non-significant correlation
(*) indicates a significant correlation with p < 0:05
(**) indicates a highly significant correlation with p < 0:01

Table 2: Cross-linguistic correlations of richness indices studied

5. Discussion

In the previous section, we have shown that certain basic stylistic regularities appear to be systematically persistent irrespective of the language studied. In less formal terms, people who send lengthy tweets in English do so in Spanish as well, and vice versa. We have also shown that people who use big words when they write in English may also use big words when they write in Spanish, and vice versa, and that people with a complex English vocabulary also have a complex Spanish one. All three of these are well-known and well-studied stylistic variables, but their cross-linguistic persistence is a novel finding with significant implications.

Similarly, people choose to engage in a particular kind of social discourse on Twitter and retain that choice across their language of participation. The same systematic effects were also shown in the use of three Twitter-specific extra-linguistic conventions. People who participate in these conventions in one language are also likely to participate in these conventions in the other. However, this is not simply a function of participation levels; the correlation between the frequency of @mentions in English and the frequency of #hashtags, also in English, by the same user, is actually negative [-0.4145, p (two-tailed) < 0.14]. This shows that there is at

best no relationship between the use of mentions and hashtags, and possibly even a slight tendency for people who use hashtags not to use mentions, and vice versa.

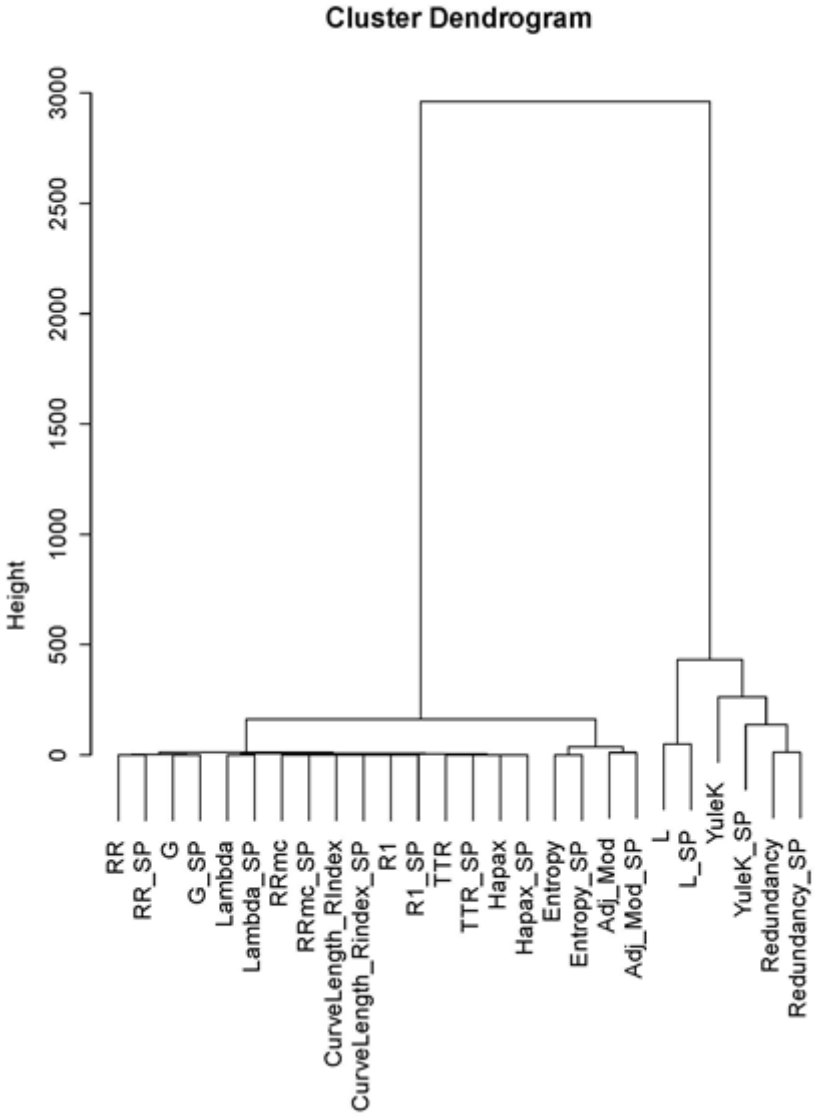

Figure 1: Cluster analysis of correlations of richness indices studied (using `truncated' measurements)

This, in turn suggests that participation in extra-linguistic conventions on social media is not dependent upon specific languages, but a personal choice that can be detected and used for inference.

In other words, cross-linguistic authorship analysis should be practical. Based on these findings, it is in principle possible to address questions like authorship of the Fuat Agni Twitter stream (Kocagul, 2014) by showing (hypothetically) that Fuat Agni uses unusually long tweets with an unusually low number of hashtags, while a particular candidate author uses short tweets and mostly posts tagged messages. Therefore, goes the hypothetical argument, that particular candidate is not the person behind Fuat Agni. This argument could be made even if the candidate author mostly tweeted in another language such as English. Similarly, by showing an unusual distribution of properties (perhaps during one month, Fuat Agni used unusually short words and unusually few hashtags, while during the following

month, the reverse was true), our finding could support a conclusion that Fuat Agni was multiple authors instead of just one.

The most obvious future work needs are replication and extensions. This work needs to be extended to types of documents other than just tweets. Candidates include other forms of social media (such as Facebook posts) but also more traditional types of writing such as novels, news articles, personal letters, and so forth. We also plan to examine other language pairs to see whether similar effects would hold in languages other than Spanish and English.

6. Conclusion

The idea of individual variation in language is not controversial, nor is the idea that certain types of variation are systematic and persistent across a single person's writings across a representative corpus. This paper has shown that these types of variation are also systematic and persistent across a highly unrepresentative set of writings as well. In our corpus, collected from people who used Twitter in both English and Spanish, we were able to show a very high correlation between ordinary stylistic variables measured on the two languages. For example, people who send long tweets do so irrespective of language. People who use a wide Twitter vocabulary in one language do so in the other language. The evidence for average word length is more ambiguous but may suggest similar cross-linguistic similarities.

In addition to these traditional stylistic variables, we have shown similar results for several extra-linguistic features related to the social conventions of Twitter itself. As with lengthy Tweets, people who post Tweets with a high (or low) number of @mentions tend to do so across languages. We have similar data for the use of #hashtags and of embedded hypertext links beginning with “http:”; for all of these conventions, the degree of participation in these conventions appears to be an individual decision rather than a language property.

This kind of data is routinely used to validate or challenge authorship of documents in a single language or genre (Juola, 2014). Grant, in particular, has used this kind of data to address issues in a murder case (Grant, 2013). The implications of this study, however, are that the same kind of data can be used across languages.

References

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance, 16*(2), 9-17.

Brooks, R. (14 July 2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed, *The Sunday Times*. Retrieved from http://www.thesundaytimes.co.uk/sto/news/uk_news/Arts/article1287513.ece

Brooks, R., & Flyn, C. (14 July 2013). JK Rowling, the cuckoo in crime novel nest, *The Sunday Times*. Retrieved from http://www.thesundaytimes.co.uk/sto/news/article1287601.ece

Burrows, J. F. (1989). „An ocean where each kind. . .“: Statistical analysis and some major determinants of literary style. *Computers and the Humanities, 23*(4), 309-321. doi: 10.1007/bf02176636

Chaski, C. E. (2005). Who“s at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence, 4*(1), 1-13.

Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy, XXI*(2), 441-466.

de Morgan, A. (1851/1882). Letter to Rev. Heald 18/08/1851. In S. Elizabeth & D. Morgan (Eds.), *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*. London: Longman‟s Green and Co.

Grant, T. (2013). Txt 4n6: Describing and measuring consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy, XXI*(2), 467-494.

Herper, M. (10 March 2014). Linguistic Analysis Says Newsweek Named The Wrong Man As Bitcoin's Creator. *Forbes*.

Hoover, D. (2004). Delta prime? *Literary and Linguistic Computing, 19*(4), 477-495.

Juola, P. (2008). Authorship attribution. *Foundations and Trends® in Information Retrieval, 1*(3), 233-334. doi: 10.1561/1500000005

Juola, P. (2013a). How a Computer Program Helped Show J.K. Rowling write A Cuckoo‟s Calling. *Scientific American*. http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

Juola, P. (2013b). Stylometry and immigration: A case study. *Journal of Law and Policy, XXI*(2), 287-298.

Juola, P. (2014). The Rowling case: A proposed standard protocol for authorship attribution. *Proceedings of Digital Humanities 2014*. Lausanne, Switzerland.

Juola, P., Noecker, J. I., Stolerman, A., Ryan, M. C., Brennan, P., & Greenstadt, R. (2013). Keyboard behavior-based authentication for security. *IT Professional, 15*, 8-11.

Kocagul, H. (2014). Authorship analysis in an anonymous Twitter user in Turkish: Who is playing the game? *Proceedings of the First African Regional Conference of the IAFL (FLFFA 2014)*. Sfax, Tunisia.

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA: Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Layton, R., Watters, P., & Dazeley, R. (2010). Authorship Attribution for Twitter in 140 Characters or Less *2nd Workshop on Cybercrime and Trustworthy Computing Workshop (CTC), 19-20 July 2010, Ballarat, Australia* (pp. 1-8).

McMenamin, G. R. (2011). Declaration of Gerald McMenamin. http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin

Mikros, G. K., & Perifanos, K. (2013). Authorship attribution in Greek tweets using multilevel author‟s n-gram profiles. In E. Hovy, V. Markman, C. H. Martell & D. Uthus (Eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California* (pp. 17-23). Palo Alto, California: AAAI Press.

Sousa-Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2011). „twazn me!!! ;(‟ Automatic Authorship Analysis of Micro-Blogging Messages. In R. Muñoz, A. Montoyo & E. Métais (Eds.), *Natural Language Processing and Information Systems* (Vol. 6716, pp. 161-168). Berlin / Heidelberg: Springer.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy, XXI*(2), 420-440.

Van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP), 4*(1), 1-17. doi: 10.1145/1187415.1187416