Using function words and punctuation marks in Arabic forensic authorship attribution

David García-Barrero, Manuel Feria and Maria Teresa Turell *

Abstract. *This paper presents an approach to written Modern Standard Arabic forensic authorship attribution drawn for the first time on a full tokenization of the text and based on the analysis of several variables such as type/token ratio, word length in characters, punctuation, conjunctions, the combination of punctuation and conjunctions, and the standard deviation of sentence length in words, among others. These variables have been tested in a sample corpus of three Moroccan writers producing two genres (short stories and literary reviews) in two measurement times. The hypotheses to be tested are: (a) There will be more inter-author than intra-author variation; (b) A writer's idiolectal style will stay stable throughout time; and (c) This idiolectal style will not be so stable when constrained by genre. The texts are segmented (tokenized) with TOKAN, part of MADA 3.2 (Morphological Analysis and Disambiguation for Arabic – CADIM group, Columbia University), based on Tim Buckwalter's Aramorph 1.2.1, which allows simplifying the detection of variables involving agglutinated clitics. Preliminary observations show that some of these variables may be discriminant markers for authorship attribution.*

Keywords: Arabic, authorship attribution, Arabic variation, inter- and intra-author variation, Arabic tokenization.

Introduction

This paper reports on the use of some quantitative techniques in authorship attribution applied to written Modern Standard Arabic (MSA). So far, several semi-automatic and quantitative approaches have been applied to authorship attribution of real-world and real forensic case texts. These approaches include linear discriminant analysis (Bel *et al.*, 2012), the use of reference corpora, Bayesian likelihood ratio methods (applied to both

*IULA, Universitat Pompeu Fabra; Departamento de Traducción e Interpretación, Universidad de Granada; and IULA, Universitat Pompeu Fabra

oral and written texts), and measurements such as lexical density analysis, among others (See Grieve 2007 and references therein for an evaluation of these techniques applied to authorship attribution.)

However, despite the prominence given to quantitative methods in languages such as English and Spanish, rather less attention has been paid to Arabic. To the best of our knowledge, previous literature on quantitative measurements applied to Arabic authorship attribution is limited to Abbasi and Chen (2005a,b, 2006); Estival *et al.* (2007) and more recently, Ouamour and Sayoud (2012) and Sayoud (2012).

With the purpose of bridging the gap between the state of the art in other languages and in Arabic authorship attribution studies, this article presents an Arabic-driven approach. This approach consists specifically in a quantitative analysis drawn, for the first time, on a full tokenization of the Arabic texts. For this purpose, all clitics have been segmented. Tokenization of clitics is of the utmost importance for two main reasons:

1. Clitics and affixes have an overwhelming statistical impact when dealing with Arabic. Using an Arabic corpus containing 600 million words, Alotaiby *et al.* (2010) showed that the corresponding lexicon size was reduced by 24.54% when applying clitic tokenization produced by AMIRA 2.0 (Habash, 2010: 89) on Arabic Gigaword (Graff, 2007).
2. Among the clitics we can find the three one-letter non-derived prepositions (Ryding, 2005: 366 and following); the ubiquitous conjunctions و *w#*[1] and ف *f#*, which are the scaffolding of the Arabic text (Warraki and Hassanein, 1994); the even more ubiquitous definite article ال *Al#*, pronouns and +k. We must take into account as well that several clitics can appear together in one word. All these function words are crucial for any quantitative approach to Arabic authorship attribution.

Table 1 shows clitics tokenized by Alotaiby *et al.* (2010) on Arabic Gigaword (Graff, 2007)[2].

On the other hand, 'there was no Western-style punctuation in Classical Arabic [...]. In general, the coordinating conjunctions and discourse markers served as punctuation [...] Modern Written Arabic has adopted[3] and adapted Western punctuation, without abandoning certain features of the Classical Arabic system (especially noticeable in coordination)'

This fact, which poses a major problem for text segmentation (Ameur *et al.*, 2008), teaching translation from Arabic into Western languages (Ghazala, 2004) and (Dendenne, 2010) and teaching English as a L2 for target learners whose L1 is Arabic (Awad, 2012) can be efficiently exploited in authorship attribution studies, especially considering punctuation and conjunctions together. This kind of analysis can give us insight into variation in discourse structure, dialectal and diachronic variation in written MSA.

For the purpose of this study, sequences such as a comma followed by a copulative conjunction, or by any other connector, can be measured in an n-gram based approach taking characters, graphic words or tokens, or parts of speech (POS) as basic units. We assume that punctuation marks and conjunctions operate in MSA at the same discursive level and define an Arabic token as 'a space delimited unit in clitic tokenized text' (Diab *et al.*, 2004). Tokenization also solves, to a great extent, the homography issue, which has an enormous impact when dealing with Arabic, as explained below.

Table 1. Clitics' Statistics according to Alotaiby *et al*. (2010)

Clitics	Transliteration	Frequency	%
ال	Al#	103,015,016	57.02
و	w#	31,014,498	17.17
ل	l#	11,470,854	6.35
ه	+h	10,021,855	5.55
ب	b#	8,857,080	4.9
ها	+hA	7,975,714	4.41
هم	+hm	2,264,578	1.25
ف	f#	1,462,196	0.81
س	s#	1,348,962	0.75
نا	+nA	1,144,143	0.63
ك	k#	737,638	0.41
هما	+hmA	449,528	0.25
ي	+y	302,246	0.17
ك	+k	267,685	0.15
ني	+ny	147,805	0.08
كم	+km	125,782	0.07
هن	+hn	35,215	0.02
ا	+A	21,342	0.01
كما	+kmA	5,611	0
كن	+kn	1,057	0

Objectives

This work is part of an on-going research project whose purposes are as follows:

1. To validate in written MSA hypotheses in authorship attribution already verified for other languages such as English, Spanish or Catalan.
2. To shed some light on which variables are potentially discriminant for Arabic authorship attribution.
3. In the long term, this research aims for the establishment of a reliable attribution methodology that makes it possible to conduct scientifically rigorous forensic text comparison in Arabic.

Hypotheses

Attribution studies rely on the notion of idiolectal style, as defined by Turell (2010). This means that each author or speaker has his/her own style that can be measured and characterizes his/her use of language. The main hypothesis that has already been tested for other languages is that:

1. For certain linguistic variables, there is more inter-author variation than intra-author variation, i.e. more variation in texts that have been written by different authors than in texts written by the same author.
   There are obviously other factors affecting variation within texts written by the same author, such as communicative purposes, register, genre and the span of time separating two samples of writing. Therefore, the above-mentioned leading hypothesis can be further elaborated into two more specific null hypotheses:

2. For the same variables, there is less variation between authors writing in the same genre (inter-author intra-genre) than between texts of two different genres written by the same author (intra-author inter-genre), and

3. For the same variables, there is less variation between authors writing within a short period of time (inter-author intra-time) than within the same author in two sufficiently separated measurement times (intra-author inter-time).

Corpus

1. In order to avoid diatopic variation, 3 Moroccan authors produced the writing samples: أحمد المديني Ahmed Al-Madini (author 1), محمد عز الدين التازي Mohamed Azzedine Tazi (author 2) and محمد برادة Mohamed Berrada (author 3). The rationale for choosing Morocco is no other than the authors' familiarity and the interest of the Spanish scholars in the country for obvious geostrategic reasons. We expect to be able to increase the number of authors in the future.
2. In order to avoid diastratic variation and context-based variation in the broad sense, the selected authors belong to the same educational level. The samples were randomly selected from comparable reputed sources, such as monographs (compilations of short stories or literary criticism articles of an author or several authors published by reputed publishers) and national newspapers (mainly the weekly cultural supplements of العلم Al-Alam and الاتحاد الاشتراكي Al-Ittihad al-Ichtiraki). Quotes (in literary criticism articles) and dialogs (in short stories), when present, were removed from the analysis.
3. In order to control inter-genre and inter-author variation, the selected texts include no quotations and the corpus was stratified according to genre into two sets: short story, a genre strongly related to narrative discourse, and literary criticism, a genre strongly related to expository-argumentative discourse. We use here the term *genre* in the traditional, literary sense (and not in the narrower sense defined by Swales 1990).
4. In order to control inter-time variation, we have included samples in two measurement times with a time lapse of nine or more years for each author and each genre.

We have collected 5 equally sized samples for each author, genre and time, which make up a total of 60 samples. Dealing with equally sized samples allows us to avoid text-length dependency problems in this stage of the research process. The sample size has been determined in approximately 650 graphic words (always including complete sentences). This Arabic-driven sample size has been calculated adapting the English 800 words standard, considered an average for forensic real-world texts, by taking into account the agglutinative character of the Arabic script in the light of the conclusions drawn by Alotaiby *et al.* (2010).

Processing

The texts were processed using MADA 3.2, a tool developed by Columbia's Arabic Dialect Modelling Group (CADIM), at Columbia University. This tool provides disambiguation, lemmatization, POS tagging and full tokenization for Arabic.

MADA is based on the information provided in BAMA (Buckwalter's Arabic Morphological Analyzer), or its latest version SAMA 3.1 (Standard Arabic Morphological Analyzer). We have used a free version of BAMA: Aramorph 1.2.1,[4] that allows MADA

to be run since its newest release (MADA 3.2) in February 2012. The free access to this tool represents a noteworthy step forward in Arabic NLP and in Arabic linguistics.

In this research, only TOKAN (the tokenization module of the MADA tool) has been used. In Table 2, we show an example of MADA's capabilities, with an input of a sequence of two Arabic *graphic words* that are split into core words and clitics in TOKAN. MADA also provides POS information on each token.

Table 2. MADA-TOKAN example

Input	وخصوصياتها wxSwSyAthA			العربية AlErbyp	
Segmentation (TOKAN)	ـها +hA	خصوصيات xSwSyAt	و w#	عربية Erbyp	الـ Al#
Part of speech	Pronoun (enclitic)	Noun	Conjunction (proclitic)	Noun	Def. article (proclitic)

Linguistic variables

Tables 3 to 9 describe the linguistic variables used in the analysis: definite article and pronouns (3), prepositions (4), conjunctions (5), demonstratives (6), negative particles (7), punctuation marks (8) and the combination of punctuation marks and conjunctions (9). Despite the fact that the size of the samples has been strictly controlled, data, with the exception of ratios, are expressed as relative frequencies (variable / total number of tokens). In addition to the frequencies of specific words, several combinations have been taken into account, namely sums of words of the same part of speech or sharing similar features, and ratios measuring the relationship between tokens (or sums of tokens) of related parts of speech or features. For instance, we have linked definite articles and enclitic pronouns (variable [6]), since they are mutually exclusive when occurring with nouns; we have also linked the substantive subordinating conjunction and the relative pronouns [47] in order to measure the substantive and relative subordinations.

The variables shown above are not all possible Arabic function words, but only those that had a considerable frequency in the text samples. For that reason, we have only considered third person singular masculine and feminine pronouns and demonstratives.

After the tokenization process, the ambiguity of the Arabic script is solved to a great extent: most of the output tokens represent disambiguated, specific words and parts of speech. However, some ambiguity still affects two important tokens due to homography:

1. The lack of short vowels affects مِن *mn* [25], which has been considered here only as a preposition ('of', 'from'), but which can also stand for the relative *wh*-pronoun ('who', 'whom'). This kind of ambiguity will be solved in the future by simply introducing short vocalization: مِن *min* (preposition), مَن *man* (pronoun). The preposition لِـ *li* [30] has also a counterpart verbal particle لَـ *la*, though it is very unlikely to be tested in MSA.
2. In a much lower degree of ambiguity, orthographic variation affects أن *>n* [46], that can stand for أنْ *>ano* (similar to the infinitive particle 'to') and أنَّ *>an a*

Table 3. Variables: Definite article and pronouns

No.	Variable	Description
1	ال	Proclitic definite article *Al#*
2	ـه	Enclitic pronoun 3rd p. sg. masc *+h*
3	ـها	Enclitic pronoun 3rd p. sg. f. *+hA*
4	ـها / ـه	Ratio enclitic pronouns 3rd p. sg. m. / f.
5	ـه + ـها	Total enclitic pronouns 3rd p. sg.
6	ال / ـه+ـها	Ratio definite article / total enclitic pronouns [1, 5]
7	هو	Independent pronoun 3rd p. sg. m. *hw*
8	هي	Independent pronoun 3rd p. sg. f. *hy*
9	هو + هي	Total independent pronouns 3rd p. sg.
10	ـه + ـها + هو + هي	Total enclitic + independent pronouns [5, 9]
11	ـه+ـها / هو+هي	Ratio enclitic / independent pronouns [5, 9]
12	ـه + هو	Total m. pronouns [2, 7]
13	ـها + هي	Total f. pronouns [3, 8]
14	ـه+هو / ـها+هي	Ratio m. / f. pronouns [12, 13]
15	الذي	Relative m. pronoun *Al∗y*
16	التي	Relative f. pronoun *Alty*
17	الذي + التي	Total relative pronouns m. + f.
18	ـه + هو + الذي	Total m. pronouns [2, 7, 15]
19	ـها + هي + التي	Total f. pronouns [3, 8, 16]
20	ـه+هو+الذي / ـها+هي+التي	Ratio m. / f. pronouns [18, 19]
21	ـه+ـها+هو+هي+الذي+التي	Total pronouns (enclitics + independents + relatives) [5, 9, 17]

('that'), both considered here as substantive subordinating conjunctions. Furthermore, the first letter أ > can be spelled as ا *A*, which is a non-normative wide-spread orthographic variant of both أ > and إ < (see Parkinson 1990 and Buckwalter 2004, and therefore MADA+TOKAN normalizes both into ا *A* in the preprocessing. Thus, in this case, أن >*n* is a homograph for إن <*n*, which itself stands for إنْ <*ino* (conditional subordinating conjunction) and إنَّ <*in a* (emphasis particle).

On the other hand, some function words have not been considered in this particular study due to their structural lexical ambiguity, such as ما *mA*, which stands for a negative particle or a relative pronoun. Interestingly, several highly ambiguous words appear combined in compound function words. Testing if discriminatory potential increases with increasing disambiguation provides an important direction for future research.

Table 4. Variables: Prepositions

No.	Variable	Description
22	بـ	Proclitic preposition *b#*
23	في	Preposition *fy*
24	مع	Preposition *mE*
25	من	Preposition *mn*
26	عن	Preposition *En*
27	على	Preposition *ElY*
28	كـ	Proclitic preposition *k#*
29	إلى	Preposition *<lY*
30	لـ	Proclitic preposition *l#*
31	إلى + لـ	Total 'dative' prepositions *<lY + l#*
32	بـ+في+مع+من+عن+على+كـ+إلى+لـ	Total prepositions [22–30]
33	$\frac{\text{إلى}}{\text{لـ}}$	Ratio 'dative' prepositions *<lY/l#*
34	$\frac{preposition}{\text{بـ+في+مع+من+عن+على+كـ+إلى+لـ}}$	Ratio preposition / total prepositions [22–31, 32]

Table 5. Variables: Conjunctions

No.	Variable	Description
35	و	Proclitic copulative conjunction *w#*
36	فـ	Proclitic copulative conjunction *f#*
37	ثم	Copulative conjunction *vm*
38	و + فـ + ثم	Total copulative conjunctions [35–37]
39	أو	Disjunctive conjunction *>w*
40	و + فـ + ثم + أو	Total coordinating conjunctions [35–37, 39]
41	$\frac{\text{و}}{\text{و+فـ+ثم}}$	Ratio *w#* / copulative conjunctions [35, 38]
42	$\frac{\text{و}}{\text{و+فـ+ثم+أو}}$	Ratio *w#* / coordinating conjunctions [35, 40]
43	$\frac{\text{فـ}}{\text{و}}$	Ratio *f#* / *w#* [35, 36]
44	$\frac{\text{فـ}}{\text{و+فـ+ثم}}$	Ratio *f#* / copulative conjunctions [36, 38]
45	$\frac{\text{فـ}}{\text{و+فـ+ثم+أو}}$	Ratio *f#* / coordinating conjunctions [36, 40]
46	أن	Subordinating conjunction *>n*
47	$\frac{\text{أن}}{\text{الذي+التي}}$	Ratio *>n* / relative pronouns [46, 17]

Table 10 shows additional descriptive data used, such as type-token ratio, standard deviation of sentence length in words and word-length distribution (i.e. frequencies of the total of words of n-characters). All these measurements have been provided automatically by WordSmith analysing both raw and tokenized texts, with the exception of variable [79] (ratio of tokens in raw and tokenized texts). Except for the n-character

Table 6. Variables: Demonstratives

No.	Variable	Description
48	هذا	Proximal demonstrative m. sg. *h∗A*
49	هذه	Proximal demonstrative f. sg. *h∗h*
50	ذلك	Distal demonstrative m. sg. *∗lk*
51	تلك	Distal demonstrative f. sg. *tlk*
52	هذا + هذه + ذلك + تلك	Total demonstratives [48–51]
53	$\frac{\text{هذا+هذه+ذلك+تلك}}{sentences}$	Ratio demonstratives / sentences [52, 81]

Table 7. Negative particles

No.	Variable	Description
54	لا	Negative particle *lA*
55	لم	Negative particle *lm*
56	لا + لم	Total negation [54, 55]
57	$\frac{\text{لا+لم}}{sentences}$	Ratio negative particles / sentences [56, 81]

Table 8. Punctuation marks

No.	Variable	Description
58	.	Single dot (stop)
59	..	Two dots (ellipsis)
60	...	Three dots (ellipsis)
61	.. + ...	Total ellipsis [59, 60]
62	. + .. + ...	Total stop + ellipsis [58, 61]
63	،	Comma
64	. + .. + ... + ،	Total stop + ellipsis + comma [62, 63]
65	$\frac{،}{.}$	Ratio comma / stop [63, 58]
66	$\frac{.}{sentences}$	Ratio stop / sentences [58, 81]
67	$\frac{،}{sentences}$	Ratio comma / sentences [63, 81]
68	¶	Newline
69	$\frac{.}{¶}$	Ratio stop / newline [58, 68]
70	$\frac{،}{¶}$	Ratio comma / newline [63, 67]
71	"	Quotation marks (double quotes)

word frequency [84], that shows significantly different data on raw and tokenized texts for obvious reasons, these measurements appear fairly stable when their relative frequencies are obtained dividing the variable in question by either the number of 'tokens' in the raw or in the tokenized texts.

Table 9. Punctuation marks and conjunctions

No.	Variable	Description
72	، و	String comma + copulative conjunction ,w#
73	$\frac{\text{، و}}{\text{،}}$	Ratio string ,w# / comma [72, 63]
74	. و	String dot + copulative conjunction ,w#
75	$\frac{\text{. و}}{\text{.}}$	Ratio string .w#/ dot [74, 58]
76	، و + . و	Total strings [72, 74]
77	$\frac{\text{، و} + \text{. و}}{\text{و}}$	Ratio strings / copulative conjunction w# [76, 35]
78	$\frac{\text{، و} + \text{. و}}{\text{.}+\text{..}+\text{...}+\text{،}}$	Ratio strings / stop + ellipsis + comma [76, 64]

Table 10. Descriptive data (WordSmith)

No.	Variable	Description
79	$\frac{tokens_{sg}}{tokens_{or}}$	Ratio tokens in segmented / original texts
80	$\frac{types}{tokens}$	Type-token ratio
81	$sentences$	Number of sentences
82	$\frac{tokens}{sentences}$	Sentence mean length in words (ratio tokens / sentences)
83	$\sigma(tokens \in sentence)$	Standard deviation of sentence length in words
84	$n{-}character\ words$	Words with n-characters (letters)

Analysis

In order to test the performance of the measurements, an analysis of variance (ANOVA) and several linear discriminant analyses (LDA) were conducted.

The degree of relation between the variables and the groups of authors, genres and times was established by means of an ANOVA test. Table 11 shows the set of ten variables most strongly related to authorship. Within these top ten variables, the most recurrent are combinations of punctuation marks and conjunctions.

Furthermore, the most discriminant variables for Arabic authorship attribution selected by the LDA are:

1. Combinations of punctuation marks and conjunctions:
   (a) Number of the copulative conjunction و w# occurring after a comma [72] or a dot [74].
   (b) Sum of the preceding variables divided by the total number of و w# [77].
2. Punctuation marks:
   (a) Number of commas divided by number of newlines (i.e. average of commas in each paragraphs) [67].
   (b) Number of stops divided by number of sentences (i.e. how much sentences end with a dot) [66].
   (c) Number of ellipses (made of two or three following dots) [61].

Table 11. The 10 variables most strongly related to authorship according to ANOVA

No.	Variable	Description	p-value
80	$\frac{types}{tokens}$	Type-token ratio	0,00000000007
72	، و	String comma + copulative conjunction ,w#	0,00000003485
63	،	Comma	0,00000008835
77	$\frac{، و+ . و}{و}$	Ratio strings / copulative conjunction w#	0,00000022231
9	هو + هي	Total independent pronouns 3rd p. sg.	0,00000028884
76	، و+ . و	Total strings	0,00000078888
73	$\frac{، و}{،}$	Ratio string ,w# / comma	0,00000174227
84	*3−character words*	Words with 3-characters (in raw texts)	0,00000750034
23	في	Preposition *fy*	0,00001525503
84	*2−character words*	Words with 2-characters (in raw texts)	0,00001540769

(d) Quotation marks [71].

3. Other copulative conjunctions: ثم *vm* [37] and the ratio of ف *fa* / و *w#* [43]. Interestingly, the former variable is a discriminating variable in text samples of literary criticism articles and the latter in text samples of short stories.
4. In addition, data obtained automatically using WordSmith also proved to be discriminatory in these analyses: the type-token ratio [80], some n-character words [84] (namely 2- and 3-character words in the raw samples) and the standard deviation of sentence length in words [83].
5. Less importantly, other function word frequencies were selected, such as:
   (a) The subordinating conjunction أن *>n* [46].
   (b) The preposition على ElY [26] and its relative frequency with respect to the rest of the prepositions [34].
   (c) The negative particle لم *lm* [55], the total of negative particles لا *lA* and لم *lm* [56] and its relative frequency with respect to the number of sentences [57].

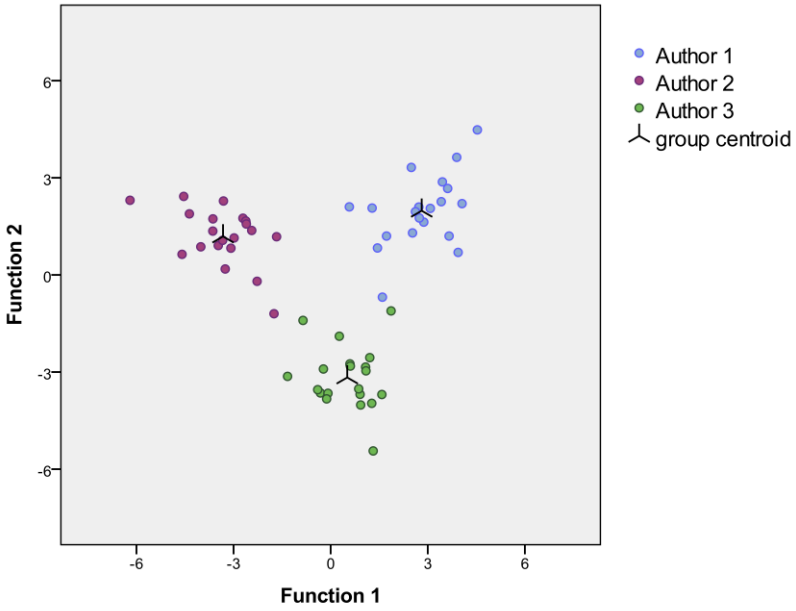
The first test classified the whole corpus by authors with a high accuracy of 59 out of 60 correct assignations using 12 variables (Figure 1). The same test obtained 56 out of 60 in the cross validation classification, in which every sample text is considered as a 'disputed' text to be assigned to an author. The variables were selected using default values of F (3,84–2,71). To gain reliability, we have modified the F values to decrease the number of variables selected, following the norm exposed by Poulsen and French (2004):

> As a "rule of thumb", the smallest sample size should be at least 20 for a few (4 or 5) predictors. The maximum number of independent variables is n - 2, where n is the sample size. While this low sample size may work, it is not encouraged, and generally it is best to have 4 or 5 times as many observations and independent variables.

A summary of the results with default and modified F values with the original and cross validation classifications for this test and the following ones is given in Table 11.

Genre is hypothesized to be another source of variation, as stated before. Therefore, the test was repeated after dividing the corpus by genre. Every single literary criticism

Figure 1. Classification by author (60 samples)



sample was correctly assigned to its author with 10 variables selected (Figure 2); only 1 sample out of 30 was misclassified in the cross validation classification after decreasing the number of variables to 5. Similar results were obtained with short story samples (Figure 3). No interference of the variable "time" was observed in either case.

Figure 2. Classification by author (30 samples of literary criticism articles)

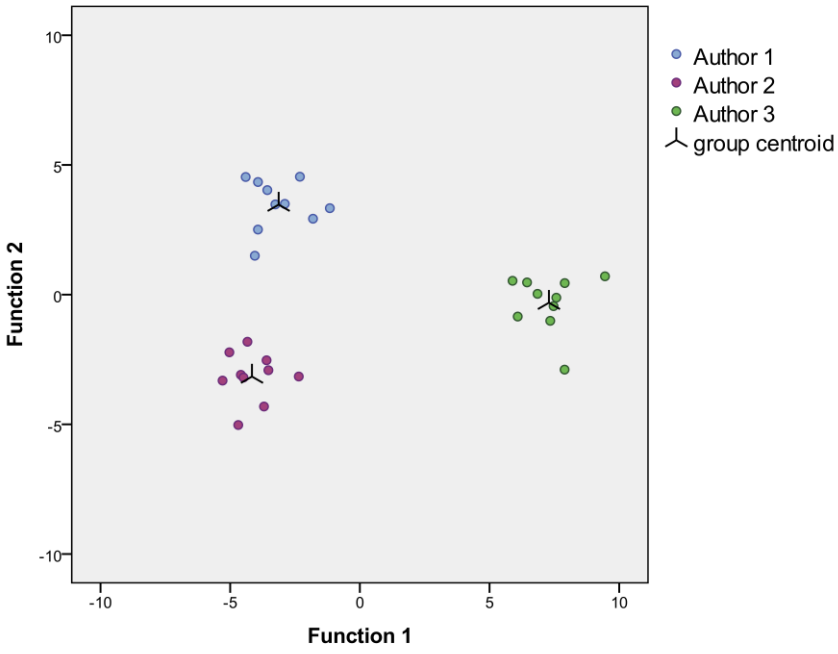


Figure 3. Classification by author (30 samples of short stories)

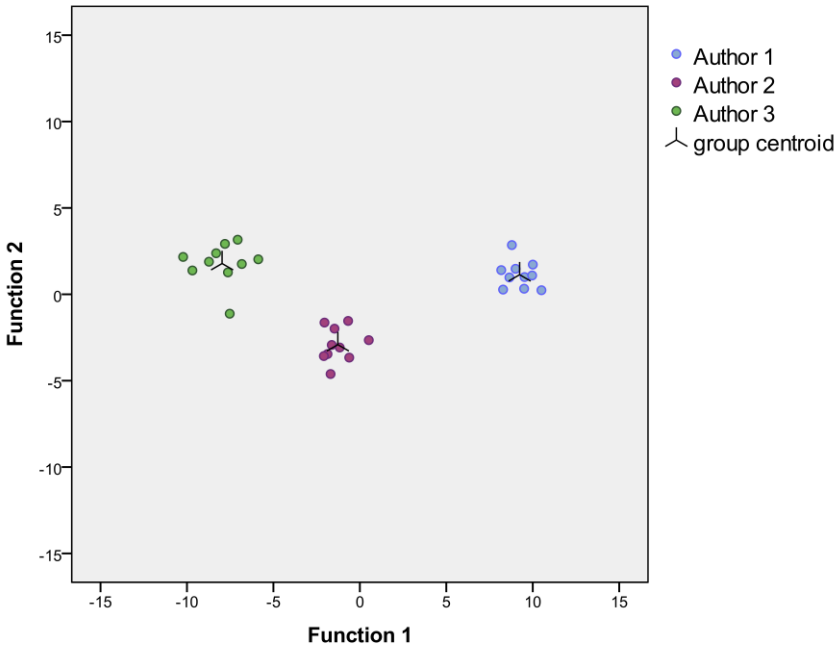


Table 12. Classification results

	F values	Variables selected	Original classification accuracy	Cross validation accuracy
60 samples	3,84-2,71 (default)	12	98,3% (1 error)	93,3% (4 errors)
	5,5-4,5	9	96,7% (2 errors)	95% (3 errors)
	7-6	7	98,3% (1 error)	95% (3 errors)
	8-7	5	93,3% (4 errors)	85% (9 errors)
30 samples of literary criticism articles	3,84-2,71 (default)	10	100%	100%
	5,5-4,5	5	100%	96,7% (1 error)
30 samples of short stories	3,84-2,71 (default)	11	100%	100%
	5,5-4,5	5	100%	96,7% (1 error)

Conclusions and future research

In conclusion, we have obtained very positive results at a not very much sophisticated disambiguation level. As for the hypotheses, we can reject our null hypotheses and consider proved that it is possible to attribute authorship using discrete variables also in Arabic texts. Genre showed to be an influential independent variable, which has to be controlled in the attribution process, whereas time showed no interference on it.

In our view, this research is a pioneer work that contributes to Arabic authorship analysis in particular, and to forensic authorship attribution in general. Moreover, it contains theoretical and methodological proposals that contribute to Arabic linguistics.

Future research will explore other measurements to cover a wider range of variables. Further advantage will be taken of different possibilities of disambiguation, such as the POS tagging provided by MADA. This would let us apply the POS n-gram approach that has successfully been applied to other languages at ForensicLab (the forensic linguistics laboratory at the Institut Universitari de Linguistica Aplicada of the Universitat Pompeu Fabra). Finally, all tests will be conducted in the future using a larger corpus of Moroccan authors and a reference corpus of writers with different national backgrounds.

Acknowledgments

This paper has been supported by the Spanish Ministry of Education and Science via its Research Project FFI2008-03583/FILO (PI: M. Teresa Turell) and the predoctoral scholarship FPU Program AP2010-5279.

Notes

[1]We use Buckwalter (2002) transliteration, except when it comes to proper names. For clitic boundaries, we use the 'tatweel' elongation character (-) for the Arabic ligated clitics (all except وw#), and in transliteration, following Alotaiby *et al.* (2010) and others, number sign () for proclitic boundary and plus sign (+) for enclitic boundary.

[2]The authors consider سـ s and ـا +A as clitics, although سـ s is rather an inflectional morpheme indicating future time in the verb, and ـا +A a case morpheme that could also be considered as a derivational morpheme turning an adjective into an adverb. Unfortunately, adding all the percentages presented by Alotaiby *et al.* (2010) the result is over 100%.

[3]Its use was first introduced in 1911 by أحمد زكي باشا Ahmed Zaki Basha in his book الترقيم وعلاماته في اللغة العربية Altrqym wElmAth fy Allgp AlErbyp (Punctuation and its Marks in Arabic Language).

[4]According to MADA developers, 'In our tokenization tests, an Aramorph MADA build reproduced the same tokenization as a SAMA MADA build for 99.4% of the words tested'. Source: `https://lists.cs.columbia.edu/pipermail/mada-users/2012-February.txt`

References

Abbasi, A. and Chen, H. (2005a). Applying authorship analysis to arabic web content. *Intelligence and Security Informatics*, 75–93.

Abbasi, A. and Chen, H. (2005b). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5), 67–75.

Abbasi, A. and Chen, H. (2006). Visualizing authorship for identification. *Intelligence and Security Informatics*, 60–71.

Alotaiby, F., Foda, S. and Alkharashi, I. (2010). Clitics in arabic language: a statistical study. *Proceedings of Pacific Asia Conference on Language, Information and Computation*, 24, 595–602.

Ameur, A. T., H., M. and W., A.-S. (2008). Semantic-Based Segmentation of Arabic Texts. *Information Technology Journal*, 7, 1009–1015.

Awad, A. (2012). The most common punctuation errors made by the English and the TEFL majors at An-Najah National University. *An-Najah University Journal for Research (Humanities)*, 26, 1.

Bel, N., Queralt, S., Spassova, M. . and Turell, M. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tombilin, N. MacLeod, R. Sousa-Silva and M. Coulthard, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham, UK: Aston University.

Buckwalter, T. (2002). Arabic transliteration.

Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. In *Proceedings of the COLING 2004 Workshop on computational approaches to Arabic script-based languages*, 31–34.

Dendenne, B. (2010). *The Translation of Arabic Conjunctions into English and the Contribution of the Punctuation Marks in the Target Language. The Case of Wa, Fa and Thumma in Modern Standard Arabic.* Dissertation submitted in partial fulfilment of the requirements for a master degree in applied language studies, Mentouri University, Constantine, Algeria.

Diab, M., Hacioglu, K. and Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. In *5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, Boston, MA.

Estival, D., Gaustad, T., Pham, S. B., Hutchinson, B. and Radford, W. (2007). Tat: an author profiling tool with application to Arabic emails. In *Proceedings of the Australasian Language Technology Workshop*, 21–30.

Ghazala, H. (2004). Stylistic-semantic and grammatical functions of punctuation in English-Arabic translation. *Babel*, 50(3), 230–245.

Graff, D. (2007). *Arabic Gigaword Third Edition.* Philadelphia: Linguistic Data Consortium.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing.* Toronto: Morgan and Claypool Publishers.

Ouamour, S. and Sayoud, H. (2012). Authorship attribution of ancient texts written by ten Arabic travellers using a SMO-SVM classifier. In *Communications and Information Technology (ICCIT), 2012 International Conference*, 44–47: IEEE.

Parkinson, D. B. (1990). *Orthographic Variation in Modern Standard Arabic: The Case of ' the Hamza*, volume Perspectives on Arabic Linguistics II: Papers from the Second Annual Symposium on Arabic Linguistics. John Benjamins Publishing Company, 72 edition.

Poulsen, J. and French, A. (2004). Discriminant function analysis (DA).

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic.* New York: Cambridge University Press.

Sayoud, H. (2012). Author discrimination between the Holy Quran and Prophet's statements. *Literary and Linguistic Computing*, 27(4), 427–444.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

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

Warraki, N. N. a. and Hassanein, A. T. (1994). *The Connectors in Modern Standard Arabic*. Cairo: The American University in Cairo Press.