Sample Size in Arabic Authorship Verification

Hossam Ahmed
Leiden University Institute for Area Studies
Witte Singel 25, 2311 BZ, Leiden, The Netherlands
h.i.a.a.ahmed@hum.leidenuniv.nl

Abstract

Authorship Verification aims at identifying whether a document of questionable authorship is created by a specific author, given a number of documents known to have been written by that author. This type of authorship analysis uses feature engineering of feature sets extracted from large documents. Given the nonlinear morphology and flexible syntax of Arabic, feature extraction in large Arabic texts requires complex preprocessing. The requirement of large training and testing documents is also impractical for domains where large documents are available in print, given the scarcity of reliable Arabic OCR. This problem is approached by investigating the effectiveness of using an author profiling-based approach on a small set of shorter documents. The findings show that it is possible to outperform the state-of-the-art authorship verification method by using a small set of training documents. It is also found that an increase in the size of the training or testing corpus does not correlate with improving the accuracy of the authorship verification method.

1 Introduction

Authorship Verification (AV) is a type of authorship analysis task where a document of questionable attribution is judged as to whether it is written by a certain author, given a number of documents known to be written by that author. AV tasks are often compared to Authorship Attribution (AA) tasks, where a document of unknown attribution is attributed to one of a number of candidate authors. AV has a number of applications in forensic linguistics and literary studies in areas where an AA task cannot answer the problem at hand. For example, while an AA task is appropriate in some cases of plagiarism detection, an AV task can better suite a situation where the text is not written by any of the candidate authors, or when there is only one candidate author.

This paper examines the effect of small sample size on the accuracy of AV tasks. Specifically, it addresses the following question: is it possible to use small testing and training datasets without significant accuracy sacrifices in an Arabic AV task? Recent developments in AV (and AA) have achieved high rates of accuracy using various Machine Learning (ML) techniques and feature configurations. Current research (c.f. section 2) achieves accurate AV results using relatively large training and testing corpora. A smaller training and/or testing set is, of course, advantageous. For one thing, a smaller data size allows for more efficient processing. For another, in real-life situations, there may not be plenty of large texts available for the AV task. Either the question document or the authentic corpus could be of small size. In a situation specific to Arabic literary studies, a great deal of documents is only available in non-machine readable format, and in typeface that does not allow for efficient OCR. Digitizing large texts for the purpose of automatic AV is, then, an unduly expensive procedure. In this paper I examine the effects of using a small corpus for training or testing documents on the accuracy of predicting AV in different domains in Modern Standard Arabic (MSA) through a number of AV experiments.

This paper is organized as follows: section 2 outlines a brief review of literature on AV, Arabic AV and AA, and how sample size is handled in the relevant literature. Section 3 describes the corpus and features used in the experiments. Section 4 describes the verification method. Section 5 describes the procedure of the two experiments conducted. Section 6 outlines the results and I

discuss their implications in 7. Section 8 is the conclusion.

2 Related Work

Statistical methods in AA have been the subject of much recent research. Ouamour and Sayoud (2013) show that ML methods (specifically SVM) perform better than purely statistical AA tasks. Howedi and Mohd (2014) and Altheneyan and Menai (2014) use naïve Bayes to test AA in Classical Arabic texts. Other ML algorithms used vary from Linear Discriminant Analysis (LDA) (Shaker, 2012) to Naive Bayes. Altakrori et al. (2018) examine a variety of ML algorithms (Naive Bayes, SVM, Decision Trees, Random Forests, and cosine distance) to examine Arabic AA in Twitter data. In terms of feature selection, successful features used include rare word unigrams (Ouamour and Sayoud, 2013), function words (Shaker, 2012), and function word and punctuation (García-Barrero et al., 2013). Altakrori et al. (2018) examine a large variety of features (character, word, and sentence counts; average word and sentence lengths, ratios of characters, short words, blank lines; punctuation, and diacritics; as well as function words).

As far as AA is concerned, a survey of Arabic AA by Ouamour and Sayoud (2018) shows that Manhattan Distance and SMO-SVM give best accuracies. It has been possible to achieve high accuracy using small text datasets. García-Barrero et al. (2013) use 650-word document samples written in MSA and Ouamour and Sayoud (2018) use 10 books (10 extracts each) of 550 average word length, achieving 90% accuracy using Manhattan Distance.

2.1 Authorship Verification in Arabic

While AA and AV share much of task characteristics, the essential difference is the lack of negative evidence. AA is essentially a classification problem, where a Question Document is put in the class of the author to which it is most similar. In AV, however, available data comes from only one author. Although this scenario is more likely to happen in real-world applications (e.g. a section of a text being added from another source), it is much more difficult to characterize and solve than an AA problem. Available data is only a corpus of work by a single author, and a single document of questionable attribution to that author (Stamatatos, 2009).

To handle the challenge of the absence of negative evidence, two approaches are generally followed (Halvani et al., 2017). In the Imposter Method, a supplementary dataset of documents not written by the same author as the authentic documents, converting the problem into an AA problem. Altakrori et al. (2018) implicitly follow this approach for Arabic Twitter posts. Although their stated scenario is that of law enforcement, they frame the AV problem as determining the true author of tweets from a list of suspects. It is likely in that context that law enforcement needs to determine the attribution of a tweet to a single individual one at a time, as the true author may not be any of the suspects. The second approach is Author Profiling. In that approach, features form documents of known authorship are extracted and used to calculate a profile of the author. The question document is then tested against that profile. If it is similar to the profile beyond a certain threshold, it is deemed authentic. Successful similarity measures in AV include Manhattan Distance (Halvani et al., 2016; Burrows, 2002), or compression-based distance (Halvani et al., 2017). Halvani et al. (ibid) note that the second approach is more computationally efficient, as only the dataset of known documents is processed. Ahmed (2017) argues that the performance of imposter-based systems relies on the selection of the supplementary dataset, which can be contentious. To determine a similarity threshold, Halvani et al., (2017; 2016) use Equal Error Rate (ERR) for English (Halvani et al., 2017) and a number of other languages (Halvani et al., 2016), ERR is a similarity value where false positives and false negatives are equal. False positives are determined from a supplementary set of negative data. For the English, Spanish, and Greek, Jankowska et al. (2014) use the area under ROC curve to determine the threshold. For Arabic, Ahmed (2018, 2017) uses a simpler Gaussian curve and dispenses with supplementary negative data altogether.

There has been limited research on Arabic AV, all of which uses author profiling techniques and datasets of varied length. Elewa (2018) examines AV of disputed Hadith (sayings of Prophet Mohammed) as related to the distribution of lexical features (token length, token-type ratio, n least/most common tokens). It uses training and

testing sets of 20 hadiths each, averaging about 150 words each. With such small text size, the author uses multivariate analysis to manually notice relations rather than Machine Learning. Ahmed (2018) uses an array of feature n-grams (tokens, stems, trilateral roots, Part-of-Speech tags, diacritics), a similarity measure based on Manhattan Distance (Burrows, 2002), and a similarity threshold based on simple probability to investigate their use in AV in Classical Arabic on a small corpus with large document sizes (11,000 – 400,000 tokens). Although the model achieves high accuracy (87.1%), the size of the training and testing documents, as well as the type of preprocessing needed to extract the best performing feature (stem bigrams) make the task computationally expensive and unsuitable for online processing. Furthermore, such huge document size in the training corpus, while may be realistic for Classical Arabic heritage work, is uncommon in modern Arabic. All the studies above are concerned with Classical Arabic. This is the first study to investigate Modern Standard Arabic genres.

3 Corpus

To test the accuracy of an AV task in Modern Standard Arabic (MSA) with small sample sizes, a corpus taken from a number of domains is compiled. Five MSA domains are selected: fiction, nonfiction, economics, politics, and opinion columns. For each domain, texts written by 10 authors are used for training and testing. Table 1 details the composition of the corpus.

Choice of the authors and text has been governed by copyright considerations, as well as the availability of a sufficient number of texts produced by the same author to allow for training and testing at different sample sizes. Whenever possible, authors coming from the same country (Egypt) have been selected to control for cross-dialectal variation.

3.1 Feature Selection

To establish a suitable baseline for evaluation, the same features used in Ahmed (2018) have been selected. It is also the highest performing approach we are aware of for Arabic AV (albeit Classical Arabic, as opposed to MSA in this experiment). Classical Arabic and Modern Standard Arabic share essentially the same grammar (syntax and morphology). However, hundreds of years of language change have contributed to a greatly expanded lexicon. Additionally, it does not follow naturally that MSA authors make the same choices when it comes to selecting among available structures (e.g. using Verb-first vs. noun-first sentence types). However, as this is the best

Author	Documents	Source
Fiction		Hindawi Foundation repository www.hindawi.org
Ali Al-Jaarim	10	
Abdul Aziz Baraka Sakin	10	
Nicola Haddaad	10	
Nawaal Al-Saadaawi	10	
Georgi Zidaan	10	
Non-fiction		
Abbas Al-Aqqaad	11	
Ismail Mazhar	10	
Salama Moussa	10	
Fouad Zakareyya	10	
Zaki Naguib Mahmoud	10	
Economics		www.almasryalyoum.com
Musbah Qutb	10	
Mohammed Abd Elaal	10	
Bissan Kassab	10	www.madamasr.com
Waad Ahmed	10	
Yumn Hamaqi	10	www.ik.ahram.org.eg
Politics		
Alaa Al-Aswani	10	www.dw.com
Wael Al-Semari	10	www.youm7.com
Danadarawy Al-Hawari	10	www.youm7.com
Belal Fadl	11	www.alaraby.co.uk
Salma Hussein	10	www.shoroukn ews.com
Columnists		www.shoroukn ews.com
Ashraf Al-Barbari	11	
Emad Eldin Hussein	10	
Fatima Ramadan	10	
Mostafa Kamel El Sayyed	10	
Sara Khorshid	10	
Total	**253**	

Table 1: Corpus used.

Domain	Avg. size
Columnists	802
Economics	820
Fiction	1,159
Nonfiction	1,108
Politics	850

Table 2: average document size per domain.

available benchmark available for Arabic, it allows for an acceptable starting point.

A secondary, yet welcome, information that this experiment can provide is identify whether an AV technique used in Classical Arabic is also applicable to MSA, which may attest to studies related to language change and historical linguistics.

Specifically, the feature set used in this paper consists of n-grams (n = 1 – 4) of the following features:

- **Token**: individual words separated by spaces. They may include proclitics and enclitics.
- **Stem**: a token without proclitics or enclitics.
- **Root**: the triliteral root from which the word is derived.
- **Diacritics**: each token is vocalized, then letter characters are removed.
- **Part of Speech**: each document is tagged for POS using MADAMIRA tagset (Pasha et al., 2014).

3.2 Preprocessing and Feature Extraction

For pre-processing, documents are downloaded as plain text (UTF-8 encoding). Fiction and non-fiction documents are downloaded as epub and converted to plain text. Front matter of each document is removed (title, author name, name and URL of the web site, etc.). Documents longer than 1,000 words are truncated. Documents consisting of fewer than 1,000 words are used in their entirety. Table 2 shows average document size per domain. For books (fiction and non-fiction), a slice of 1,000 words is taken from the middle of each book. This decision is taken to avoid the possibility of repeated sections typical of a given author across works (for example, a repeated preface in non-fiction, or list of characters in a work of fiction). White spaces are normalized to single space, and punctuation marks are removed.

For feature extraction, tokens are defined as strings of characters separated by space. Roots, POS tags, and diacritics are generated using MADAMIRA version 2.1 with default settings. MADAMIRA output files are processed using Regular Expressions to extract relevant features to separate plain-text files.

4 Method

The purpose of this paper is to determine the effect of document size on the accuracy of AV tasks. To do so, two experiments are carried out. The first experiment uses the dataset in its entirety to determine which specific feature n-gram ensembles yield best results (i.e. highest accuracy) for each of the five domains. This experiment is motivated by the fact that the feature set used in Ahmed (2018) is tested in Classical Arabic, and should not be taken for granted that the same feature configuration will perform equally well in MSA, or similarly across genres. The second experiment uses the best performing feature for each domain and examines the change in AV accuracy with progressively smaller training set size. Linear regression analysis of the results of each experiment is conducted to estimate whether there is correlation between document or corpus size and accuracy.

4.1 Verification Method

Each verification task is divided into a number of problems. Each problem consists of a question document and a set of known documents.

In the training step, the known documents are used to calculate a similarity threshold. In the testing step, similarity between the question document and the training set is calculated. The question document is deemed authentic if its similarity value is higher than the threshold. The verification method is similar to that used in Ahmed (2018), with the difference that the current experiment uses the entire set of features, not only the most frequent n%.

4.2 Training, testing, and evaluation

For each domain, input to the training phase is a set of strings representing the feature in question known to be attributed to a given author. N-grams of appropriate value for n are generated using NLTK (Bird et al., 2009), and relative (normalized) frequencies of the features described in the section

Feature Selection are calculated, also using NLTK. Output of the training phase is a similarity value threshold for an authentic document.

Similarity is calculated using Manhattan Distance between a document X and a corpus of known documents Y:

$$dist(X,Y) = \sum_{j=1}^{n} |X_j - Y_j| \quad (1)$$

where X_j and Y_j are the normalized frequencies of feature j. Distance is then converted into a similarity score:

$$Sim(X,Y) = \frac{1}{dist(X,Y)} \quad (2)$$

Similarity Threshold θ is calculated by determining Sim for each document in the training set in relation to the reset of the training documents, creating a confidence interval for all the training documents. θ is then calculated as the upper bound of the interval at $p<0.005$.

Testing and evaluation are done by calculating *Sim* for each test document. Accuracy is calculated as the number of correct answers divided by the total number of documents tested.

Although the aim of this paper is to evaluate the effectiveness of using different sample sizes, a task that essentially does not require a baseline, an accuracy of 87.1% will be used as a guiding baseline. This accuracy is the best accuracy achieved in the relevant literature (Ahmed, 2018), albeit coming from a different register (MSA).

5 Experiments

5.1 Experiment 1: Best performing ensembles

In order to be able to plot AV accuracies against document size, it is necessary to identify best performing feature ensemble (feature + ngram). Although previous literature (Ahmed, 2018) suggests that stem bigrams are the most successful feature combinations, it should not be taken for granted that the feature combination that has been successful for Classical Arabic is also the best performer across domains in MSA.

To select the best performing feature-n-gram ensemble for each domain, the AV task described in the previous section is implemented on the full size of the corpus. For each domain, the accuracy of each feature ensemble is evaluated using the leave-one-out method. Table 3 shows the best performing feature combination for each domain.

Domain	Features	Accuracy
Columnists	Stem bigrams	80%
	Token unigrams	80%
	Diacritic unigram	80%
Economics	Root bigrams	76.8%
Fiction	Diacritic bigrams	84%
Nonfiction	Stem unigrams	81.57%
Politics	Token unigrams	84.53%

Table 3: Best performing feature ensemble per domain.

The results of experiment 1 show that with a test document size averaging 850 – 1000 tokens, best performing features vary by MSA domain. None of the domains achieved an accuracy close to the baseline, although the two domains that score lowest accuracy (economics and columnists) have the lowest document average size.

5.2 Experiment 2: Document Size Effects

There are three factors in play for determining size effects in AV: the size of the question document, the number of training documents, and the size of the training set overall. Experiment 2 examines all three variables.

For Experiment 2, the training and testing procedure for Experiment 1 is replicated 6 times, using only the highest performing features as indicated in Experiment 1, and with varying sizes of the training set $S \in \{5, 6, 7, 8, 9, 10\}$ documents. The result of the experiment is an ordered set (Q, T, R), where Q is the size of the question document, T is the size of the combined training set, and R (1, 0) is the result of the verification process. R = 1 if the correct prediction is made, and R = 0 if an incorrect prediction is made. Accuracy is calculated for values of Q in intervals demarked by $Q \in \{0, 500, 600, 700, 800, 900, 1000, 1100, 1200\}$, and for $T \in \{0, 5000, 6000, 7000, 7500, 8000, 8500, 9000, 11000\}$. Two datapoints are excluded (fiction T = (8000, 8500) and nonfiction T = (7000, 7500)) as outliers. Each of the two datapoints consist of one document and have R = 0%. Linear regression analysis between accuracy and relevant size variable is then conducted using SPSS.

Domain	Features	Training set (documents)	Accuracy
Columnists	Stem bigrams	5	**87.84%**
		6	87.45%
		7	85.10%
		8	84.71%
		9	81.57%
		10	80.00%
Economics	Root bigrams	5	**90.00%**
		6	88.00%
		7	86.00%
		8	84.00%
		9	81.20%
		10	78.40%
Fiction	Diacritic bigrams	5	**89.20%**
		6	86.80%
		7	84.80%
		8	84.00%
		9	83.00%
		10	84.00%
Nonfiction	Stem unigrams	5	**89.80%**
		6	87.84%
		7	87.45%
		8	84.71%
		9	83.14%
		10	81.57%
Politics	Token unigrams	5	**90.59%**
		6	87.06%
		7	85.49%
		8	85.49%
		9	83.92%
		10	83.53%
Baseline	**Stem bigrams**	**19**	**87.1%**

Table 4: Best performing feature ensemble per domain.

6 Results

Table 4 shows the results of testing the verification method using the leave-one-out method on a training corpus of 5 – 10 documents, and using the best-performance features identified in Experiment 1. In all five domains, the verification method performs best at S = 5 training documents. Regression Analysis shows a strong correlation coefficient of -0.931, with $p<0.005$ (c.f. Figure 1).

Regression analysis to identify correlation between the accuracy of the verification method and the total size of the training set in tokens is conducted. As Figure 2 shows, there is a moderate positive correlation of 0.492, with $p<0.05$ ($p = 0.003$).

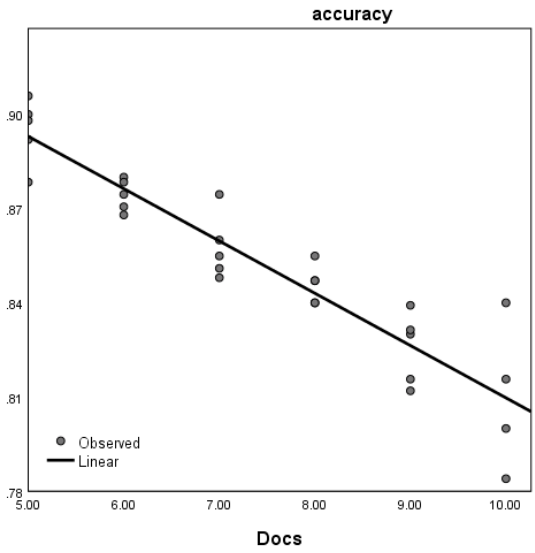

Figure 1: Number of training docs – accuracy correlation.

Regression analysis between accuracy of the verification method and the size of the test document in tokens does not show any significant correlation between the two variables (coefficient of correlation = 0.132, p = 0.48).

7 Discussion

The results of Experiment 2 show that a training set with a smaller number of documents outperforms one with a larger number of documents.

In every domain, a training set of five authentic documents outperforms the baseline of 87.1% in classical Arabic. This finding is consistent with Altakrori et al.'s (2018) observation for AA that fewer candidate authors generally contribute to better performance. The finding in this paper extends the scope of that statement to Modern Standard Arabic AV number of training documents.

Another informative finding of the experiment is the lack of significant correlation between the size of the question document and AV task performance. The implication of this finding in Digital Humanities and literary studies is that if the suspect document is an entire book, there is no need to digitize the whole document. This is especially useful for Arabic given the vast amount of print resources, and lack of reliable affordable OCR.

A rather unexpected result from Experiment 2 is the positive correlation between training set size in

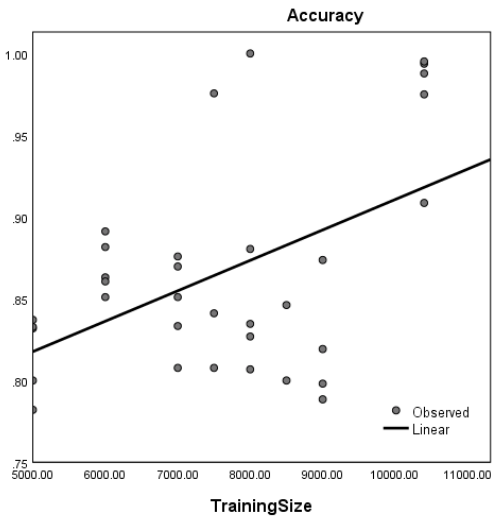

Figure 2: Training set size – accuracy correlation.

tokens on the one hand and accuracy on the other. If there is significant strong negative correlation between the number of documents in the training set, one would expect negative or no correlation between the sum of those training documents and accuracy. One could rule this result out as coincidence, but this is unlikely, given a low p-value (0.003). A possible explanation for this result could be the difference in average document size of the fiction and nonfiction corpora. The higher average document size in these two domains means that all the observations related to those two domains are clustered towards the upper bound of the word count, including their highest accuracy observations; the results with fewer training documents (e.g. S = 5) for fiction and non-fiction are in the same band for larger S for other domains. Indeed, this seems to be the case. When the regression analysis is repeated excluding measurements for fiction and non-fiction, regression for the remaining three genres show no statistical significance.

8 Conclusion and future work

This paper shows that high AV accuracy can be achieved using relatively small sample size for the training corpus (5 documents). It also shows that for document size < 1000 words, having a larger training or testing sample does not affect the performance of AV. The findings of this paper are of particular interest in the context of literary and journalistic analysis.

There are a number of areas that future research can cover. First, this paper shows that smaller training sets result in improved accuracy, when applied to the set of features that perform best on experiment 1 (full training set). Future research can investigate if other feature ensembles can outperform the ones tested in experiment 2, but were not considered here because of steeper degradation in accuracy at training set size S = 10. The accuracies reported here rely in part on the accuracy of feature extraction as well as on the distance measure used (Delta, (Burrows, 2002)). The accuracy of the feature extraction using MADAMIRA is around 96%, depending on the feature extracted (Pasha et al., 2014). As better morphological analyzers ae developed, future research should consider the effects of better feature extraction on the selection of features to be used. Additionally, other distance measures should be considered, in addition to Manhattan Distance. Finally, It is unclear if high AV accuracy based on this method can be achieved in other domains where document sizes are necessarily shorter, such as online product reviews and social media communications. Nonlinguistic features such as punctuation and non-Arabic characters were also not investigated. I leave these questions for future research.

References

Hossam Ahmed. 2017. Dynamic Similarity Threshold in Authorship Verification: Evidence from Classical Arabic. *Procedia Computer Science*, 117:145–152.

Hossam Ahmed. 2018. The Role of Linguistic Feature Categories in Authorship Verification. *Procedia Computer Science*, 142:214–221.

Malik H Altakrori, Benjamin C M Fung, Steven H H Ding, Abdallah Tubaishat, M H Altakrori, S H H Ding, and Farkhund Iqbal. 2018. Arabic Authorship Attribution: An Extensive Study on Twitter Posts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, 18(1):51.

Alaa Saleh Altheneyan and Mohamed El Bachir Menai. 2014. Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University - Computer and Information Sciences*, 26(4):473–484.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

John Burrows. 2002. "Delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–

287.

Abdelhamid Elewa. 2018. Authorship verification of disputed Hadiths in Sahih al-Bukhari and Muslim. *Digital Scholarship in the Humanities*.

David García-Barrero, Manuel Feria, and Maria Teresa Turell. 2013. Using function words and punctuation marks in Arabic forensic authorship attribution. In Rui Sousa-Silva, Rita Faria, Núria Gavaldà, and Belinda Maia, editors, *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*, pages 42–56, Porto, Portugal. Faculdade de Letras da Universidade do Porto.

Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship verification for different languages, genres and topics. *Digital Investigation*, 16:S33–S43.

Oren Halvani, Christian Winter, and Lukas Graner. 2017. Authorship Verification based on Compression-Models.

Fatma Howedi and Masnizah Mohd. 2014. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. *Computer Engineering and Intelligent Systems*, 5(4):48–56.

Magdalena Jankowska, Evangelos Milios, and Vlado Kešelj. 2014. Author Verification Using Common N-Gram Profiles of Text Documents. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*:387–397.

Siham Ouamour and Halim Sayoud. 2013. Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features. In *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 144–147.

Siham Ouamour and Halim Sayoud. 2018. A Comparative Survey of Authorship Attribution on Short Arabic Texts.

Arfath Pasha, Mohamed Al-badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M Roth. 2014. MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*:1094–1101.

Kareem Shaker. 2012. *Investigating features and techniques for Arabic authoriship attribution*. Ph.D. thesis, Heriot-Watt University.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.