Authorship Attribution Revisited: The Problem of Flash Fiction
A morphological-based Linguistic Stylometry Approach

Abdulfattah Omar
Department of English, College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia
&
Department of English, Faculty of Arts, Port Said University, Port Said, Egypt

Basheer Ibrahim Elghayesh
Department of English, Faculty of Languages and Translation,
Al Azhar University, Cairo, Egypt
&
Department of English, College of Science and Arts,
Sajer, Shaqraa University, Saudi Arabia

Mohamed Ali Mohamed Kassem
Department of English, College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia

Abstract
This study is concerned with addressing the limitations with the authorship attribution of flash or micro-fiction. The shortness of linguistic data in texts of the kind makes it challenging for conventional stylometric authorship methods to assign disputed texts to their real authors. As thus, this study proposes a new stylometric authorship system based on morphological patterns and letter mapping properties. The assumption is that these carry unique and distinctive stylistic features that can be usefully used to recognize possible authors of disputed texts. The study is based on a corpus of 259 flash fiction stories written in Arabic. Cluster analysis was for grouping documents that have shared linguistic features together. Results indicate that all texts were successfully matched with their real authors. It can be concluded that morphological information can be usefully used for improving the performance of authorship attribution and detection in Arabic texts due to the unique stylistic features of the affixation processes in Arabic. Controversial texts in Arabic can thus be assigned to their authors based on detecting stable morphological patterns with reliable authorship performance.

1. Introduction

The recent years have witnessed an increasing use of linguistic stylometric approaches in addressing different authorship problems. These have been mainly based on the investigation of the lexical (e.g. frequency of distinctive words, discourse markers, and modal verbs) and structural (e.g. use of chunks, type of sentence, and sentence length) properties of the texts as a clue for identifying authors of controversial texts. In spite of the success of these approaches in solving different authorship problems of various historical documents and literary texts, so far they are ineffective and thus unreliable in addressing authorship problems with the very short texts including what came to be known in the literature as flash fiction.

Flash fiction, Galef (2016) argues, is currently used as an umbrella or catchall term for any minuscule narrative. Initially, narratives of the kind were described as 'short-shorts', then, it came to be known as 'sudden fiction' or 'micro-fiction' with the publication of Robert Shapard and James Thomas' *Sudden Fiction* in 1986 where narratives used to range between 250-500 words. The growing popularity of this new kind of fiction has recently encouraged many authors to become more interested in producing even smaller texts , so they are read by more people (Botha, 2016). It is no surprise then that these days, some narratives are written in just two sentences described as 'Twitter fiction' or 'Twiction'(Crum, 2017). The brevity of this type of fiction has made it more accessible through social media platforms, and consequently, more authorship issues are emerging. The unique stylistic nature of these narratives makes it difficult for standard linguistic stylometry approaches to address such problems.

Technically speaking, the shortness of linguistic data in such texts provides no sufficient clues and creates problems of sparsity or data sparseness which make it challenging for conventional stylometric approaches to identify the authors of disputed texts. In the face of the limitations of existing linguistic stylometry methods, this study proposes the use of morphemes instead of lexicons and sentence structures as inputs for the linguistic stylometry of the texts. The assumption then is that the way words are combined can be useful for recognizing possible authors of disputed texts. By way of illustration, this study is based on a corpus of 259 flash fiction stories written in Arabic. The rationale is that very few studies have been done to Arabic usually applying standard stylometry approaches with no regard to the different linguistic system of Arabic. Unlike English and different Western languages, the way words are formed and built or what can be described as the morphological structure of words represents one of the unique stylistic features in Arabic. Many nouns, for instance, have more than one plural form. The word أخ [ax] meaning brother, for example, has four plural forms. All of them are morphologically valid and customarily used. An investigation of the morphological properties of words can be used thus as a clue for attributing disputed texts to their authors.

The research questions of this study are thus asked about the effectiveness of the use of the morphological structures and patterns as well as the way words are built in Arabic as variables in authorship tasks. The hypothesis is that the use of derivational and inflectional morphemes represents one of the unique stylistic features for Arab writers. The use of inflectional and derivational morphemes in Arabic can be considered a stylistic potential and powerful expressive means that can distinguish authors. Unfortunately, morphological structures have been ignored in the stylometric authorship applications in Arabic. These have always been based on standard authorship processes with no regard of the peculiar nature of Arabic morphology. In such methods, affixes, which carry rich stylistic features in Arabic, are usually removed using stemming in

standard classification applications as natural language processing (NLP) systems are designed to reduce the number of forms of words to be stored. In so doing, NLP systems often do not include any morphological processes. In light of this argument, this study asks the following research questions. First, are morphological structures and patterns useful in improving the authorship detection of Arabic texts? Second, is it possible to suggest an alternative authorship system that considers the peculiar nature of Arabic morphology?

2. Literature review: authorship attribution

Authorship attribution, also called authorship recognition, is the process of looking for salient features in a piece of writing that relates the work to its author. Craig (2004) points out that "authorship studies aim at 'yes or no' resolutions to existing problems, and avoid perceptible features if possible, working at the base strata of language where imitation or deliberate variation can be ruled out" (2004, p. 273). The idea of authorship attribution is very old. Love (2002) says that it "reaches back as far as the great library of Alexandria and embraces the formation of the Jewish and Christian biblical canons".(p. 1). The motive behind authorship attribution studies is that many works were written anonymously and many others raise suspicion about their real author, and historical evidence is sparse or lacking. Traditionally, work on authorship attribution was conceived as an organized scholarly enterprise where it was not "the work of a specialist in authorship but of a scholar for whom the determination of authorship has repeatedly been a crucial element in other kinds of investigation" (Love, 2002, p. 1). There are many examples where the task of identifying the author of a particular document was the job of politicians, journalists, and lawyers (Juola, 2008; Juola, Sofko, & Brennan, 2006). Studies in this tradition often used criteria for relating works to authors on chronological and epistemological bases. One problem with such methods is that it is often difficult to find reliable historical facts or knowledge-based evidence that will help in the identification of authors. Furthermore, studies based on what can be considered philological did not use replicable methods, and therefore, the results were not objective and thus unreliable.

In the face of these limitations, empirically-driven approaches for authorship attribution problems were developed. The claim was that authorship attribution applications should be algorithmically processed without any reference to existing analytical results or personal knowledge of authors (Moisl, 2009). The mainstream of these approaches is described in the literature as stylometry. Stylometry is a quantitative investigation into the characteristics of an author's style. Laan (1995) defines the term as a technique "to grasp the often elusive character of an author's style, or at least part of it, by quantifying some of its features" (p. 271). Similarly, Merriam and Matthews (1994) indicate that "stylometry attempts to capture quantitatively the essence of an individual's use of language." (p. 203). Stylometric studies have been mainly based on computational and quantitative methods to reach solid conclusions regarding the authorship of a given text (Tambouratzis & Vassiliou, 2007). Accordingly, numerous studies have come to provide empirical solutions to different controversial authorship issues using quantitative methods for investigating the stylistic and linguistic properties of authors.

One of the pioneering examples of the use of stylometric analysis in authorship problems is Mosteller and Wallace (1964) attempt to give internal evidence for the authors of the disputed Federalist Papers based on linguistic and stylistic properties of the authors. These are 77 Federalist Papers written during 1787-1788 to Alexander Hamilton, John Jay, and James Madison. These papers were published in newspapers under the pseudonym of *Publius until they* were collected

with eight more articles to form a volume. There was a consensus about the authorship of these Papers that John Jay had authored five papers in the volume; while Hamilton authored fifty-one papers; Madison wrote 14 and both Madison and Hamilton co-authored three. The authorship of 12 papers in the volume was somewhat disputed since it was challenging to find out which of the two, Madison or Hamilton had authored those Papers (Rudman, 2012; Savoy, 2013). On their part, Mosteller and Wallace (1964), employed tools of statistical analysis to investigate the mystery of authorship of the Federalist papers in the early 1960s, using function words as discriminators. The objectivity and replicability of the proposed approach opened the way to the digital age of authorship attribution. In literature, different studies have come to adopt stylometric methods for resolving some of the controversial authorship issues that have long been considered unanswerable. One of the typical examples of authorship attribution is the investigation of Shakespeare's plays. The main question they addressed was: Did Shakespeare write all of his plays? These studies tended to investigate whether Shakespeare's plays were written by Shakespeare himself, collaboratively with other authors, or entirely with other authors (Craig & Kinney, 2009; Erne, 2008; Hoover, 2002). The majority of these studies focused on the Marlowe-Shakespeare debate and this can be attributed to the similarity between the two authors.

The underlying assumption behind stylometric testing of authorship attribution is that, Holmes (1998) contends, "authors have an unconscious aspect of their style, a style which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive" (p. 111) and the identification of such personal distinctive linguistic and stylistic features makes it possible to detect an author's signature and distinguish the writing of one author from another or others. In this way, researchers and particularly statisticians, Knaap and Grootjen (2007) argue, have tended to investigate the lexical features of texts to make predictions about possible authors. According to (Burrows, 2003, 2007), the search for the most frequent words has been one of the most widely used methods for determining the author of a given work. Garcia and Martin (2007) explain that statisticians attempted over the last decade to solve some controversial authorship problems by finding a formula grounded on the computation of tokens, word-types, and most frequently-used words. They contend that computational statisticians have tended to investigate what they call the 'Lexical Richness' of authors to propose a reliable approach to authorship attribution. In turn, Morton (1986) argues that the use of rare words is a good indication for determining the author of a given text as this enables one writer to be distinguished from another. He explains:

> The once occurring words convey many of the elements thought to show excellence in writing, the range of a writer's interests, the precision of his observation, the imaginative power of his comparisons, they demonstrate his command of rhythm and of alternations".(p.1)

Similarly, Blatt (2017) asserts that rare words are quite noticeable and can be considered writer's favorite words which makes it easier and accurate to use them as an indicator for determining authors.

The ineffectiveness of the lexical representation of texts in resolving different authorship problems, however, has led to the development of new methods. The lexical representation of texts has come to be known today as the traditional way of doing authorship attribution. It has been criticized for its ineffectiveness in providing solutions for the practical applications of authorship

attribution (Stamatatos, 2009; Tamboli & Prasad, 2013). The claim is that isolated or single words are not enough for assigning disputed texts to their possible writers. The idea is simply that single words are not enough to capture the structure of documents. Different studies, therefore, have been more concerned with the morphological, syntactic, and structural features of texts (e.g. morphologically complex words, use of function words, sentence length, compounding, and punctuation). In spite of the reasonable success of the newly developed methodologies in providing answers for many authorship problems, verifying the authorship of very short texts as in the case of very short stories still represents a real challenge for the practical applications of author identification. Additionally, very few studies have been concerned with authorship attribution in Arabic, where differences in language systems represent further challenges. This study tends to address this gap in the literature by proposing more reliable methods for the authorship problems concerning short stories and Arabic.

3. Methodology

To test the proposed system, this study is based on a corpus of 259 flash fiction stories written by four authors: Gamal AL-Gezery, Essam Al-Sherif, Huda Kafarnah, and Haifa Hammouda. The selected stories are published in four collections of short fiction stories entitled "فكر بنفسك" /fakkir binafsik/ (Think of Yourself), "علم أسـود" /alam ?aswad/ (A Black Flag), "تذكرة للمسـافات" /taðkarah lilmasäfat/ (A Ticket For Far Destinations), and "شــاهد حنين" /ʃahid ħanīn/ (A Witness of Yearning). Documents were represented using the vector space model (VSM). The reason is that it is conceptually simple as well as it is convenient for computing semantic similarity within documents. A data Matrix M_{ij} was built in which rows H_i represent the documents and columns M_j the morphological type variables, and the value at the M_{ij} is the frequency of lexical type j in document i. The data matrix M_{ij} was built representing the selected 259 flash fiction stories. The texts were given name codes (serialized from M001 to M259) for identification. Each matrix row, therefore, represents a lexical frequency profile for the corresponding text.

For identifying the groups with common linguistic features, cluster analysis methods are used. Cluster analysis is widely acknowledged as a successful technique for organizing any unorganized set of documents (Moisl, 2015). It is an exploratory multivariate technique for systematically finding relatively homogeneous clusters of cases based on proximity measures without prior assumptions about differences within sets of data investigated (Fielding, 2007; Kaufman & Rousseeuw, 1990; Manning, Raghavan, & Schütze, 2008). It is a deterministic process that identifies discrete categories under any inherent structure in the data (Anderberg, 1973; Everitt, 1993; Everitt, Landau, & Leese, 2001; Hair, 2006; Milligan, 1996; Punj & Stewart, 1983). It is thus an inductive technique that explicitly attempts to group data sets into discrete classes (Adams, 2003; Mirkin, 2005). The aim of cluster analysis can be summarized as grouping a collection of objects into subsets where members of each subgroup are more closely related to one another than members assigned to the other group/s. Groups are technically called clusters. Given a corpus of 259 documents, these can be clustered where members of each cluster share specific characteristics. In authorship recognition applications, the assumption is that texts grouped together are more likely to be written by the same author. To perform cluster analysis, Euclidian distance, being a straightforward measure, is used. Euclidian distance is the most widely used and is reported to provide reliable results in general. As for the clustering method, Ward linkage is used. The rationale is that the Ward linkage clustering (or what is usually referred to as increase

in sum of squares) with Euclidean measure seems to be the most convenient for the present case because it makes the clearest partitioning of the matrix rows.

4. Results

In order for the proposed system in assigning texts to their real authors to be evaluated, two processes were carried out. First, similar texts were grouped together, assuming that texts grouped together are more likely to be written by the same author. Second, clustering structures were compared to the bibliographic information of each author. To compute the similarity between texts and group similar texts together, the Ward linkage clustering method with Euclidean distance measure was used. As a result, the matrix rows are assigned to four groups. One advantage of this clustering is that it offers a solution for a traditional problem in cluster analysis-the decision of the optimal number of clusters that fits a dataset. The strong tendency towards left-branching that is associated with other clustering methods is avoided with Ward clustering. The matrix rows are assigned into four main groups, as shown in Figure 1.
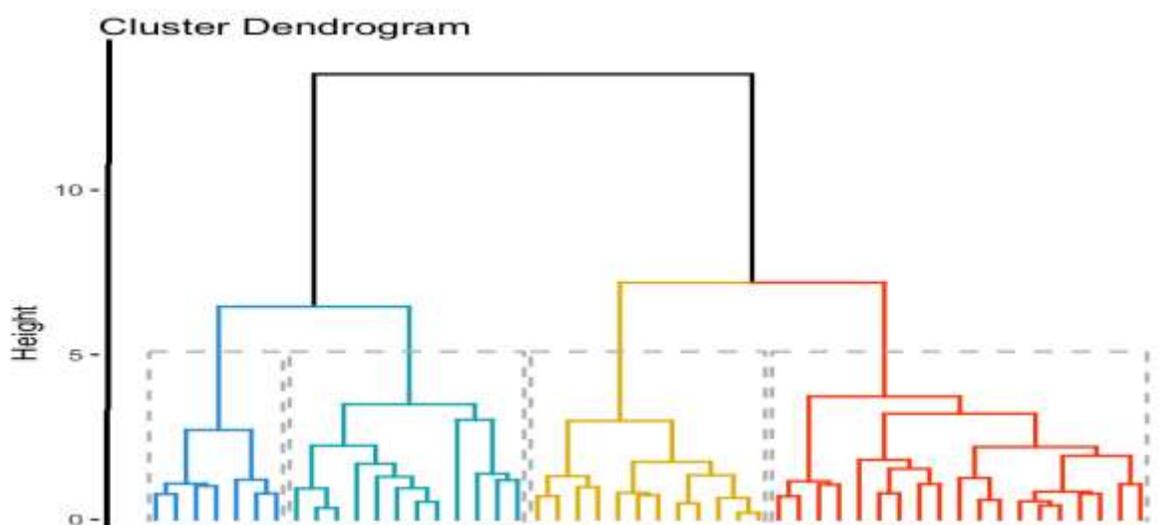

Figure 1 Cluster analysis of the selected flash fiction stories

For clustering validity purposes, two approaches were used. These are cross-validation and relative comparison. The objective is to validate the previous analysis by seeing whether the same analytical methods applied to an alternative representation of the data gives identical or at least similar results. In a cross-validation approach, the texts were randomly divided into two subsets, say *A* and *B*, and the cluster analysis is carried out separately on each of *A* and *B*. Similarity of the results is the indication of validity (Rencher, 2002). The comparison shows a close fit between the results as there is a complete correspondence between the structures based on the data matrix composed of all the 259 rows and the structures based on the random distribution of these 259 rows into two groups.

For relative comparison analysis, a comparable approach was based on comparing the clustering structure, generated by the same algorithms but using an alternative representation of the data; this was done by cluster analyzing a principal component reduction of the data matrix. The analysis showed that there is a close fit between the two clustering structures despite the minor differences. Consequently, it can be claimed that an agreement between the two clustering

structures supports the validity of hierarchical cluster analysis results. As a final step, the clustering structure obtained here, as shown in Figure 1 was compared to the bibliographic information of each author. Members of each class or cluster were compared to the stories of each author. Results indicate that all texts were successfully matched with their real authors. Members in Class A correspond to the stories written by Gamal AL-Gezery, Class B to Essam Al-Sherif, Class C to Huda Kafarnah, and finally Class D to Haifa Hammouda.

5. Analysis and Discussions

Given that the texts were clustered based on morphological similarity, this implies that each cluster has a characteristic morphological frequency profile which distinguishes it from the others. Based on this assumption, it should be possible to identify the most important and distinct variables for each group, and, based on the morphological patterns of these items, to infer the morphological characteristics of the respective groups.

Group A

In Group 1, texts have a multitude use of verbs as one of the main lexical tools used to build up the flash fictions. Al-Gezeiry, for instance, uses verbs recursively to indicate action and flow of events as well as to show the dramatic anxiety. In his flash fiction "قلة حيلة" /qillat ħīlah/ (*Helplessness*), he uses verbs such as " أشـجع – غضـب – أعطت – سـمعت" /sami3at - ?a3ŧat - ɣaɖiba - ?uʃage3/ (she heard – she gave – he got angry – I encourage). All of these verbs are tense homogeneous, i.e. used in the past. He uses inflectional suffixes to denote gender, e.g. سـمعت /sami3at/ (she heard) and أعطت /?a3ŧat/ (she gave). Besides, he uses negation tools such as "ليس" /laysa/ (not) to affirm his argument and establish a logical homogeneity. It is worth noting that all these verbs are not augmented, i.e. they do not have auxiliary consonants. The root verb is used only with the feminizing 'ta' to denote that the speaker/addressee is a female.

This use of verbs is repeatedly used in Al-Gezeiry's texts. In his "فكربنفسك" /fakkir binafsik/ (*Think of Yourself*), the author uses homogeneous verbs, such as " فكر – كتم – ساعد – استدان" /istadäna – sä3ada – katama – fakkara/ (he borrowed – he helped – he muted – he thought). Also, he uses auxiliary consonants to form augmented verbs as in "استدان" /istadäna/ (he borrowed) from the root "دان" /däna/ to indicate exaggeration in action. The use of such augmented verbs is repeated more than once in his works. For instance, in his "فراغ" /faräɣ/ (emptiness), he says "استغفر" /astaɣfiru/ which is derived from the root verb "غفر" /ɣafara/ (seek forgiveness----forgive, respectively). The use of the augmented form here denotes the request of forgiveness on the part of the speaker or the subject of the verb. Al-Gezeiry, also, uses internal gemination in his "فكَّر بنفسك" /fakkir binafsik/ (*Think of Yourself*) as in the verb "فكَّر" /fakkir/ (he thought) which is formed of two letters uttered as one. This is repeated in his "عصـر" /3aṣr/ (Afternoon), where he uses the verbs "تعمّدت، صلّيته" /ta3ammadtu/, /ṣallaytuhu/ (I intended – I prayed it). The use of augmented verbs of that form, i.e. doubled or geminated, is to denote either multitude or exaggeration. He may use both, i.e. auxiliary consonants and gemination, as in his "نبض تصـفيق" /nabɖu taṣfīq/ (the beat of a clap) in the verb "سـيصـفِّق" /sayuṣaffiq/ (he will clap). Also, the particle 'sa' before the verb 'يصـفق'/yuṣaffiq/ is to express the future tense. The use of inflectional suffixes is quite observable in Al-Gezeiry's works. For instance, in his "سـعي" /sa3y/ (pursuit), he uses the verbs "بحثت – تفحصـت – رمت" /ramat – tafaħaṣat – baħaθat/ (she saw – she inspected – she searched), where all of them are suffixed with feminizing ta (ت). It makes a kind of sound homogeneity, both in sound and tense, among them.

It can be also observed that Al-Gezeiry uses a particular verb more than once, either in the same story or in different ones. For instance, he uses the verb "صارع" /şära3a/ (he wrestled) more than once in his "شكر" /ʃukr/ (*Thankfulness*). He repeats the verb as in "يصـارعني الوميض وأصـارعه" /yuşäri3unil-wamīđu wa ?uşäri3uhu/ (the flash wrestles me and I struggle it back) which indicates a lexical and sound homogeneity. The same verb is repeated in his "ندية – ابتسـامة جانبية" /nadiyyah - ?ibtisämatun Ʒänibiyyah/ (dewiness, side smile). However, he uses the verb in different tenses, i.e. once in the past and others in the present. However, in some cases, he resorts to nouns. He starts some of his fictions either with nouns or pronouns. He may use nouns to indicate confirmation, statement or description of a specific case or event. This is obvious is his "دهاليز قصـيدة" /dahälīzu qaşīdah/ (corridors of a poem) where he uses personal pronouns, especially singular ones, e.g. " أنا – أنتَ – أنتِ – هو – هي" /?ana - ?anta - ?anti – huwa – hiya/ (I – you – He – She). He may use nouns as in "انحراف مسار" /?inħiräfu masär/ (Deflection) for the same purpose, i.e. description. He says لوحة قزحية – نهر هادر" /lawħatun qazzħiyyah – nahrun hädir/ (a rainbow painting, a babbling river). It is repeated in a number of his other works.

In some instances, Al-Gezeiry resorts to the use of vernacular or slang words. For example, in his fiction "زفت" /zift/ (tar), he says "زفت يزفتك" /zift yizaffitak/ (be tarred) where he uses the verb "يزفتك" /yizaffitak/ which is slang derived from the noun "زفت" /zift/. He applies the rules of affixation to it. He uses the object suffix "ك"; a connected pronoun, as if the original root is "زَفَتَ" /zafata/. He, too, added the prefix 'yi' to the beginning of the verb to denote third person present, i.e. imperfect indicative case. He, also, makes use of formal connecting words such as relatives and demonstratives. For instance, in his "وميض المجاهل" /wamīđul-maƷähil/ (flash of the unknown), he says "كيف دخل **ذلك** الكاتب مجاهلي **التي** لا يعرفها بشـــر؟" /kayfa daXala ðälikal-kätibu maƷähilil-latī lä ya3rifuha baʃar/ (how did that writer enter my unknown that no one knows) where he uses the demonstrative "ذلك" /ðälika/ and the relative "التي" /?allati/ to connect ideas and establish an internal lexical homogeneity.

Group B

Texts in this group are characterized by the frequent use of verbs. Al-Sherif uses verb structures very often where nominal phrases could express the same meaning. For instance, in his "عتاب" /3itäb/ (reprimand), he uses the verbs "صـمتت – قالت – أحبتك – مدّت – تمسـح" /şamatat – qälat - ?aħabbatka – maddat – tamsaħ/ (she was silent, she said, she loved you, she extended, she wipes). He, also, makes use of suffixes such as feminizing ta as in his "إضاءة" /?iđä?ah/ (lighting) in verbs such as "تسـاءلت – نظرت – وجدته" /tasä?alat – nadžarat – waƷadathu/ (she wondered – she looked – she found him). The use of augmented verb 'تسـائل' is to serve exaggeration on the part of the subject, i.e. not merely a straightforward question. This is repeated in many of his works such as دافع – معرفة قديمـة – دعاء" /däfi3 – ma3rifah qadīmah – du3äa?/ (a motive, old acquaintance, a prayer) and many others. However, he may still resort to nouns to start his fictions or internal sentences as in his "حصـــيلـة" /ħaşīlah/ (outcome), he says "الطفلـة التي تسـولـت" /?aŧ-tiflatul-latī tasawwalat/ (the girl who begged). Here he, not only uses nouns but also relatives. A specific word may be used more than once in a number of his works. For instance, the word "معتوه" /ma3tūh/ (nuts) is repeated in "مطالب – دفاع – هذيان – عدالة" /maŧälib – difä3 – haðayän – 3adälah/ (demands, defense, delirium, justice) and many others.

The use of Arabized words is observable in some of Al-Sherif's works. For instance, in his مدرسة فيسبوكية" /madrasah fisbokiyyah/ (a Facebook school). Al-Sherif borrows the English word (Facebook), then Arabicizes it and finally derive an adjective of it (فيسبوكية) /fisbokiyyah/. This is

repeated in his "فيس بوك" /fis buk/ (Facebook) where he reuses the borrowed Arabicized word Facebook as well as other Arabicized words such as "بروفايل – بوست – لايك" /brofayil – bust – layik/ (profile, post, like). He, also, uses interrogation to serve confirmation or maybe denial and exclamation from the addressee as in his "متى" /matä/ (When). He says "متى :تسـألني: متى تنام؟ قلت /tas?alunī matä tanäm qultu mata-stayqadžtu/ (she asked me: when do you sleep? I said: when did I wake up?). This is repeatedly used in a number of his fictions such as "سـؤال – دعوى" /su?äl – da3wa/ (a question, a claim).

Group C

On her part, Kafarnah resorts to the use of an adverbial phrase or noun phrase to start a number of her flash fictions. For instance, in her "ابتسامة" /?ibtisämah/ (smile), she says "بروح تواقة للفرح..." /birawħin tawwäqatin lil-faraħ/ (with a spirit eager to be happy). The same is true in her "انتظار" /?intidžär/ (waiting), she says "...من رمال انتظارك" /min rimälin-tidžärik/ (from the sands of waiting for you). Also, in her "تعب – ضياع" /ta3ab – đayä3/ (fatigue, loss) and others, she follows in the same style. However, sometimes, she uses uncommon passive participle. For instance, in her "انتظار" /?intidžär/ (waiting), she says "وخلفتني بين سـاعاتي **كسـيرة**" /wa Xallafatnī bayna sa3ätī kasīrah/ (and she left me broken with my clocks) where the passive participle "كسـيرة" /kasīrah/ (broken) is used instead of the commonly used one; "مكسـورة" /maksūrah/ on the meter "مفعول" /maf3ūl/. She, also, may use foreign or borrowed words. In her "محتال صـغير" /muħtälun ṣaɣīr/ (a young crook), she uses the word "ماما" /mäma/ which is not Arabic, Arabicized to mean "أمي" /?ummi/ (my mother).

Group D

A common feature of the texts in this group is that Hammouda begins with nouns when she describes or makes a statement. For instance, in her "يأس" /ya?s/ (despair), she says "نفسـي التي ظننت /nafsī ?allati džanantu/ (I myself who thought) , and her "بارق أمل" /bäriqu ?amal/ (glimmer of hope) she uses no verbs at all. Yet, the use of verbs dominates most of her works. For instance, in her "لوعة" /lawa3ah/ (grief), she uses the verbs "أخبروها – سـيعود – خـالفت – تعطرت – حلّ – يعد" /?aXbirūha – saya3ūd – Xälafat – ta3ᵵarat – ħalla – ya3id/ (tell her, he will come back, she breached, she wore a perfume, he settled, he promises). She, also, uses interrogative style as in her "طعنة" /ᵵa3nah/ (a stab). She says "هل ماتت لعبتي؟" /hal mätat lu3batī/ (had my toy died) and "وهل نحن على قيد الحياة؟" /wa hal naħnu 3alä qaydil-ħayäh/ (are we alive?).

The morphological analysis of each group indicates clearly that each author has a unique way of forming and building up and forming the words. The letter information/mapping and morphological structures/patterns of each author can thus serve as distinctive linguistic stylometric features that can be usefully used for determining the authors of very short texts in Arabic, as seen in the case of the selected flash fiction stories chosen for this study. Furthermore, given the accuracy of the proposed system, we can argue that linguistic stylometric studies (including authorship systems) should consider the peculiar nature of Arabic morphology.

6. Conclusion

To address the problems with authorship attribution in very short texts in Arabic and more specifically flash fiction, this study proposed a linguistic stylometric approach that considers the morphological patterns and word combinations. Results indicate clearly that morphological

information can be usefully used for improving the performance of authorship attribution and detection in Arabic texts due to the unique stylistic features of the affixation processes in Arabic. Controversial texts in Arabic can thus be assigned to their authors based on detecting stable morphological patterns with reliable authorship performance. Although the proposed system was tested only on literary texts written in Standard Arabic, the implications of the study can be usefully used for the authorship problems in other text genres including emails, newsgroup messages, Facebook posts, and tweets as well as different Arabic varieties which still represent a real challenge for the practical applications of author identification.

About authors
Abdulfattah Omar finished his PhD in computational linguistics at Newcastle University in 2010. His research interests include digital humanities, discourse analysis, and translation studies. He is currently an Assistant Professor of linguistics in Prince Sattam Bin Abdulaziz University. Previously he worked for Newcastle University, Northumbria University, and Port Said University. ORCID: https://orcid.org/0000-0002-3618-1750

Basheer Ibrahim Elghayesh is an Assistant Professor of linguistics and translation studies at Department of English, Faculty of Languages and Translation, Al Azhar University in Egypt. His research interests focus on translation theory, pragmatics, and discourse studies.
ORCID: https://orcid.org/0000-0003-4504-0121

Mohamed Ali Mohamed Kassem is an associate professor of Applied Linguistics & TEFL. His major areas of research are writing instruction, integrating MALL in EFL teaching, vocabulary acquisition and EFL teacher education. ORCID: https://orcid.org/0000-0002-8613-0580

References
Adams, R. (2003). *Perceptions of Innovations: Exploring and Developing Innovation Classification.* (PhD), Cranfield University,
Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York; London: Academic Press.
Blatt, B. (2017). *Nabokov's Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writin*: Simon & Schuster.
Botha, M. (2016). Microfiction In Ann-Marie Einhaus (Ed.), *The Cambridge Companion to the English Short Story*. Cambridge: Cambridge University Press.
Burrows, J. F. (2003). Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York. *Computers and the Humanities, 37*(1), 5-32.
Burrows, J. F. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Lit Linguist Computing, 22*(1), 27-47. DOI:10.1093/llc/fqi067
Craig, H. (2004). Stylistic Analysis and Authorship Studies. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 273-288). Oxford: Blackwell.
Craig, H., & Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.
Crum, M. (2017, Dec 06, 2017). Twitter Fiction Reveals The Power Of Very, Very Short Stories. *The Huffington Post*.

Erne, L. (2008). Reconsidering Shakespearean Authorship *Shakespeare Studies, 36*, 26-37.

Everitt, B. (1993). *Cluster Analysis* (3rd ed.). London: E. Arnold.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed. / Brian S. Everitt, Sabine Landau, Morven Leese. ed.). London: Arnold; New York: Oxford University Press.

Fielding, A. (2007). *Cluster and Classification Techniques for the Biosciences*. Cambridge, UK; New York: Cambridge University Press.

Galef, D. (2016). *Brevity: A Flash Fiction Handbook*. New York: Columbia University Press.

Garcia, A. M., & Martin, J. C. (2007). Function Words in Authorship Attribution Studies. *Lit Linguist Computing, 22*(1), 49-66. DOI:10.1093/llc/fql048

Hair, J. F. (2006). *Multivariate Data Analysis* (6th ed. ed.). Upper Saddle River, N.J. ; London: Prentice Hall PTR.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*, 111-117.

Hoover, D. L. (2002). Frequent Eord Sequences and Statistical Stylistic. *Literary and Linguistic Computing, 17*, 157-180.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends R in Information Retrieval, 1*(3), 233-334.

Juola, P., Sofko, J., & Brennan, P. (2006). A Prototype for Authorship Attribution Studies. *Lit Linguist Computing, 21 (2)*(2), 169-178. DOI:10.1093/llc/fql019

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley& Sons, INC.

Knaap, L. v. d., & Grootjen, F. A. (2007). *Author identification in chatlogs using formal concept analysis*. Paper presented at the Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC2007), Utrecht, The Netherlands, November 2007.

Laan, N. M. (1995). Stylometry and Method. The Case of Euripides. *Lit Linguist Computing, 10*(4), 271-278. DOI:10.1093/llc/10.4.271

Love, H. (2002). *Attributing Authorship : An Introduction*. Cambridge: Cambridge University Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Merriam, T., & Matthews, R. (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic Computing, 9*(1), 1-6.

Milligan, G. W. (1996). Clustering Validation: Results and Implications for Applied Analyses. In P. Arabie, Hubert, L.J. and De Soete, G (Ed.), *Classification and Clustering*. River Edge, NJ: World Scientific Publishing Co Pte Ltd.

Mirkin, B. (2005). *Clustering for Data Mining: A Data Recovery Approach*: Taylor & Francis Group, LLC.

Moisl, H. (2009). Using electronic corpora in historical dialectology research. In M. Dossena & R. Lass (Eds.), *Studies in English and European Historical Dialectology* (pp. 68-90.). Brussels; Frankfurt: Peter Lang.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics*. Berlin, Munich, Boston: Walter de Gruyter.

Morton, A. Q. (1986). Once. A Test of Authorship Based on Words which are not Repeated in the Sample. *Lit Linguist Computing, 1*(1), 1-8. DOI:10.1093/llc/1.1.1

Mosteller, F., & Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison-Wesley Pub. Co.

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research, 20*(2), 134-148.

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (Second Edition ed.): John Wiley & Sons, INC.

Rudman, J. (2012). The Twelve Disputed 'Federalist' Papers: A Case for Collaboration. *Proceedings Digital Humanities*, 353–356.

Savoy, J. (2013, November 1-6, 2013). *The Federalist Papers revisited: A collaborative attribution scheme.* Paper presented at the Proceedings of the Association for Information Science and Technology, Montreal, Quebec, Canada.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of American Society for Information Science and Technology, 60*(3), 538-556.

Tamboli, M. S., & Prasad, R. S. (2013). Authorship Analysis and Identification Techniques: A Review. *International Journal of Computer Applications, 77*(16), 11-15

Tambouratzis, G., & Vassiliou, M. (2007). Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments. *Lit Linguist Computing, 22*(2), 207-224. DOI:10.1093/llc/fqm003