

A Computational Linguistic Approach for the Identification of Translator Stylometry using Arabic-English Text

Heba El-Fiqi

School of Engineering and
Information Technology
University of New South Wales
ADFA campus
Canberra, Australia
H.El-Fiqi@student.adfa.edu.au

Eleni Petraki

Faculty of Arts and Design
University of Canberra
Canberra, Australia
Eleni.Petraki@canberra.edu.au

Hussein A. Abbass

School of Engineering and
Information Technology
University of New South Wales
ADFA campus
Canberra, Australia
H.Abbass@adfa.edu.au

Abstract- Translator Stylometry is a small but growing area of research in computational linguistics. Despite the research proliferation on the wider research field of authorship attribution using computational linguistics techniques, the translator stylometry problem is more challenging and there is no sufficient literature on the topic. Some authors even claimed that this problem does not have a solution; a claim we will challenge in this paper. We present an innovative set of translator stylometric features that can be used as signatures to detect and identify translators. The features are based on the concept of network motifs: small graph local substructures which have been used successfully in characterizing global network dynamics. The text is transformed into a network, where words become nodes and their adjacencies in a sentence are represented through links. Motifs of size 3 are then extracted from this network and their distribution is used as a signature for the corresponding translator.

We then investigate the impact of sample size, method of normalization and imbalance dataset on classification accuracy. We also adopt the Fuzzy Lattice Reasoning Classifier (FLR) among others, where FLR achieved the best performance with a classification accuracy reaching the 70% mark.

Keywords-component; Translator Stylometry; Authorship Attributions; Network Motifs; Decision Tree Analysis; Fuzzy Classifier; Computational linguistics; Arabic-English Corpus.

I. INTRODUCTION

Identifying the author of a text is an important area of research in “*Computational Linguistics*” [1, 2]. There are many studies in “*Authorship Attributions*” [1-14]. These studies use the stylometric features of the authors to identify the original authors. These stylometric features include lexical, character, syntactic, semantic, and application-specific features [1-14].

On the one hand, the problem of authorship attributions is a difficult one and research challenges remain to exist in this area. On the other hand, the sub-problem of translator attributions is even harder and no solution for it exists so far.

Translation is a fascinating topic. While the original writer of a text had a specific mental picture in her mind and an intended message to be communicated through the text, a translator faces a different type of challenge. Successful

translation necessitates that the translator needs to form the same mental picture as the original author of the text. Good translation does not stop at the level of mapping words, but extends to mapping meaning, mental pictures, imagination, and feelings. This is called the “loyalty” dilemma, where a wide discussion in the literature exists on the importance of maintaining the spirit of the original work.

If we compare author attributions to translator attributions, we find that the former is expected to have more signatures or discriminatory factors representing the choices made by the authors. Authors have many more degrees of freedom, where they can build their own identity as authors. Translators have less. Being constrained with the original text is a non-trivial limitation. This feature alone makes translator attributions a more difficult problem than author attributions. Nevertheless, we conjecture that translators attempt to have their own touch, signatures that can be used to detect who translated what. This is the hypothesis we hold in this paper.

The problem of how to identify translator stylometry is under-studied in the literature; probably because it is a harder problem. Some argue that translated work is considered as the original author’s literature work rather than the translator’s own work; however, no one can ignore the fact that translators are individuals [15]; they make personal choices which can affect the translation process. This is what we call “*Translator Stylometry*”.

In this paper, we are going to introduce a new method that uses network motifs to identify the difference in translator’s style. Our hypothesis is reformulated as “network motifs can be used to differentiate between different translators based on their own writing stylometrics”. To test this hypothesis, we represent our datasets as networks. This is done by generating a word adjacency network for each piece of work by a translator in the dataset. To analyze and compare two networks, we can use their global statistical features; these include Shortest-path length, global centrality, clustering coefficient, etc., or their structural design principles like the network motifs. Network motifs which are initially introduced by Ron Milo et al. [16] are patterns

(represented by small subgraphs) that are repeated in a real network more often than randomly generated networks.

Our method is an instance based approach, in which each sample generates its own network representing a single piece of work. After extracting the motifs, we use a classifier to classify the translated works. The accuracy of this classifier is used to test our hypothesis on a dataset comprising 30 chapters from the Holy Quran. We focus this paper on a text written in Arabic and translated into English.

II. BACKGROUND

A. Stylometry

The translator has to make many personal decisions while he is translating a piece of work. Examples of these decisions include the choice of words, discourse markers, modal verb selection, length of sentences, frames, and their own understanding of the original text.

In 2010, Winters discusses how a translator's attitude influences her translation [17]. He uses two German translations of the original novel *"The Beautiful and Damned"* (1922) written by F. Scott Fitzgerald. He shows that different translators' views affect the macro level of the novel, and how this from his point of view extended to influence the readers' attitude as well.

Xiumei uses relevance theory to explain the translator's style [18]. The findings demonstrated that, while the translator tries to balance between the original author's communicative intentions and the target reader's cognitive environment, she is still influenced by her own preferences and abilities; the outcome of all of that introduces her style.

The sample of papers reviewed above and others [17-19] demonstrate that studies in linguistics highlight that there are translator stylistic features existing in translated texts. However, the area of automatic identification and feature extraction of translator stylistic features has not seen an equivalent breed of research. The only attempt that we are aware of was in 2001 by Mikhailov and Villikka [15]. They tried to find *"stylistic fingerprints"* for translator by extracting three lexical features as: "Vocabulary richness", "Most frequent words" and "Favorite words". Their experiment was done on Russian fiction texts in addition to their Finnish translations. The lexical features that they used in the research failed to find stylistic fingerprints for different translators. Their conclusion was summed up in their title; that it is not possible to differentiate between translators. While this conclusion is inconsistent with traditional linguistic studies, it seems that it was sufficient to turn away researchers from this line of research over the last 10 years and since the publication of Mikhailov and Villikka paper in 2001.

B. Network analysis:

Newmann [20] defined a *Network* as *"a collection of points joined together in pairs by lines"*; these points are referred to as vertices or nodes and the lines are referred to as edges. Networks are everywhere; almost any system can be represented as a network. The traditional definition of a system is that it is a group of components interacting together for a purpose; this is a

definition whereby network representation is paramount (components are nodes and interactions are through links). Many tools exist in the literature for analysing networks [21]. These tools vary from mathematical, computational, to statistical tools.

Network mining is the process whereby a problem is represented through networks and is searched to identify patterns. Examples of network mining problems include: link prediction, link type prediction, discovery of communities of interest, and discovery of infrequent or unusual patterns. Networks can be analyzed through measurements on the global and local level. Local measures attempt to capture from local constructs the global features of the network. A widely used local measure is network motifs. They are used to uncover network structural design principles [22]. Network motifs have been successfully used by different researchers in Biology [23, 24], Game theory [25], Electronic circuits [26], and Software [27].

C. Fuzzy classifier:

In our research, we are going to employ two fuzzy classifiers; the first one is *Unordered Fuzzy Rule Induction Algorithm* called FURIA [28], which is an extension to the well-known RIPPER algorithm [29] but using fuzzy rather than conventional rules. This algorithm is introduced in 2009 by Hühn and demonstrated promising results in some types of classification problems [28, 30].

The other fuzzy classifier is the Fuzzy Lattice Reasoning Classifier (FLR), which is initially introduced by Athanasiadis in 2003 [31]. This classifier uses the Fuzzy Lattices to create a Reasoning Environment. The rules are induced in a mathematical lattice data domain. FLR has the advantage that it can be incremental. It is also able to deal with missing data [32].

III. DATA

In our work, we are going to work on translation from Arabic to English. We choose the Arabic language as it is spoken by approximately 280 million native speakers [33]. It is also the official language for 26 countries¹; it is the third most official language in the world after English and French. The Arabic language is gaining interest due to its socio-political importance. Millions of Muslims (both Arab and non-Arab) are studying the Arabic Language as the Holy Qur'an (the main religious text of Islam) is written in Arabic; also, all the Islamic terms are in Arabic. Therefore, the interpretation of the Arabic language has wider significance.

In our work, we are going to use "The Holy Qur'an" and the translation of its meaning as our corpus. The Holy Qur'an is a widely popular text as it is read on daily basis by Muslims. The total number of Muslims worldwide is more than 1.3 billions [34] but not all Muslims can read Arabic. Hence, there is a need to translate Qur'an meanings accurately that reflect and respect the original meaning. Despite the existence of many translations of this text, there is considerable dispute about the loss in the

¹ As in World Atlas website <http://www.worldatlas.com/> on 28th of January 2011

translation of Qur'an meanings due to the uniqueness of its textual characteristics. Another important consideration for the choice of this text was the expectation that given its religious significance, there would be minimal difference in the translations.

We obtained our corpus data from tanzil.net² website. This website introduces 14 different English translations for the meanings of Holy Qur'an. We choose two of them; the first one is by Muhammad Marmaduke Pickthall and the second one is by Abdullah Yusuf Ali. These two translations are among the few translations that have been assessed by Khaleel Mohammed in 2005 [35] as being among the most accurate translations.

IV. METHODOLOGY

A. Network formation

To establish the word-adjacency network from the dataset, we worked on Ayah (verse) level. Each word is represented by a node, and each ordered word adjacency is represented by edge going from the first occurring word to the following word. The frequency of two words adjacency is counted and represented by edges labels. The edges here represent an "occurring- before" binary relationship.

B. Features identification

Working on the lexical level, there are different features that can be extracted like word n-gram, vocabulary richness, word frequencies, and token-based [36]. We are trying to find the linkage between the words and how frequently they are used by different translators. We also attempt to extract the frequency of occurrence of patterns of ordered words –known in linguistics as 'lexical chunks'- in the text.

C. Motif extraction

The motifs are small subgraphs, usually 3, 4 or 5 nodes. For a subgraph with three connected nodes, we have only 13 distinguished possible subgraphs as shown in Fig. 1. While for four connected nodes, we have 199 distinguished possible subgraphs, and 9364 possible subgraphs for five nodes. This study uses 3-node motifs.

To illustrate how we can extract these 13 motifs, we give an example using a sample translation by "Yousif Ali"² for chapter 112 in the Holy Qura'n in Fig. 1. The sample text is "Say: He is Allah, the One and Only; (1) Allah, the Eternal, Absolute; (2) He begetteth not, nor is He begotten; (3) And there is none like unto Him. (4)".

For example, motif 7 (M7) represent the relationship between three nodes; two way relationship between the left and upper nodes, and one way relationship between the right and upper nodes. The first relationship is represented by the ordered appearance of words "is" and "He" in Aya (3), and the other direction where "He" before "is" in Aya (1). The second relationship is represented by the ordered appearance of words "nor" and "He" in Aya (3).

In motif 8 (M8), the first relationship that appeared in motif 7 is the same, while we has another two way relationship between the upper and right nodes, represented in Aya (3) where the word "He" appeared once before "not" and the second time after it.

We scanned the network for all of these possible subgraphs and counted each of them. All matches are counted, even if there are overlapping counts. We used *MAVisto*[37] for pre-processing.

M1		Aya (1): "Say: He is Allah , the One and Only"	
M2		Aya (1): "Say: He is Allah , the One and Only ;"	
M3		Aya (3): " He begetteth not, nor is He begotten;"	
M4		Aya (1): "Say: He is Allah , the One and Only ;"	
M5		Aya(1): "Say: He is Allah , the One and Only". Aya(3): "He begetteth not, nor is He begotten;"	
M6		Aya(2): " Allah , the Eternal , Absolute	
M7		Aya(1): "Say: He is Allah, the One and Only" Aya(3): "He begetteth not, nor is He begotten;"	
M8		Aya(1): "Say: He is Allah, the One and Only" Aya(3): " He begetteth not, nor is He begotten;"	
M9		Aya(3): "He begetteth not, nor is He begotten;" Aya(1): "Say: He is Allah, the One and Only;" Aya(4): " And there is none like unto Him"	
M10		Aya(3): "He begetteth not, nor is He begotten;" Aya(1): "Say: He is Allah, the One and Only;" Aya(4): " And there is none like unto Him"	
M11		Aya(3): "He begetteth not, nor is He begotten;" Aya(1): " Say: He is Allah, the One and Only;" Aya(4): " And there is none like unto Him"	
M12		Aya(3): " He begetteth not, nor is He begotten;" Aya(1): "Say: He is Allah, the One and Only;"	
M13		Not applicable for this sample text	

Figure 1. All possible 3-nodes connected Subgraph

² *Tanzil* is a quranic project launched in early 2007 to produce a highly verified Unicode quran text to be used in quranic websites and applications.
www.tanzil.net

D. Randomization:

To randomize the network, a random local rewiring algorithm is used. This algorithm keeps the degrees of the vertices constant. This is done by reshuffling the links: If A is connected to B ($A \rightarrow B$), and C is connected to D ($C \rightarrow D$).

Then, it makes a link from A to D and from C to B instead of the old links. But before applying the new links, it checks if these links already exist in the network. If so, this process is skipped and the algorithm attempts to find other links. This check is necessary to prevent having multiple links connecting the same vertices [38]. This process is repeated several times in excess of the total number of edges in the system to generate a randomized network [39]. We considered 500 randomized networks for each sample when conducting our experiments.

E. Significance test

To calculate Z-score, we need to calculate the average and standard deviation of occurrences of a motif in all randomized networks. The z-score is calculated as the difference between the frequency of this motif in the target network and the mean frequency of the generated randomized networks divided by the standard deviation of the frequency values for these randomized networks.

Since we are testing for confidence level %95, the z normal range is from -1.96 to +1.96. If the z-score is within this range, it is not significant. If it is outside this range, it is significant.

V. EXPERIMENTS

A. Research questions

The main question we attempt to answer in this paper is: Can we use network motifs to detect translator's style? The sub-questions of our research include:

- How can data normalization affect the bias of the sample?
- What is an appropriate sample size for detecting translator stylometry?
- How can class distribution balance/imbalance affect the results?

B. Expiemneal design

For all tests we used 30 parallel samples for two translators in addition to their original texts, so that we have 30 chapters from the Qur'an with their corresponding translations for two translators.

First: to address the normalization problem, we conduct 5 different tests considering the parameters as in Table 1: In the first test, we consider absolute values: The attributes of motifs represent $f(x_i)$ and $f(y_i)$. So, we have 13 attributes representing the 13 motifs and one attribute representing class label for 40 samples for training and 20 samples for testing considering class balance. The second test, using the same configuration of the first test but we tried to minimise the translator bias to her own writing; thus, we used the attributes of motifs to represent $f(x_i)/\Sigma(f(x_i))$ for the first translator, and $f(y_i)/\Sigma(f(y_i))$. In the third Test: The attributes of motifs represent $f(x_i)/f(z_i)$ and $f(y_i)/f(z_i)$. The conjecture behind this is to minimise the bias regarding the original text. While in the fourth test; we used the

attributes of motifs to represent $f(x)/(\Sigma(f(x_i) * f(z_i)))$ as a conjecture for minimising both the original text bias and translator bias. For tests 3 and 4, we need to divide by the frequency of Arabic motifs, which is sometimes zero, so we increment the number of motifs for Arabic text by 1. In the fifth test, we used 16 attributes which are the 13 motifs in addition to the number of nodes and edges and one attribute for class label.

Second: to address the sample size problem, we repeated the previous experiments with 10 sample chapters for each translator for training and 10 samples for testing, and compared these results to their corresponding results from the first experiment. The comparison is between an experiment with 10 samples as the training size and the other with 20 samples for the training size. In both cases, we maintained the test sample constant to have a fair comparison. The test size is 10 sample chapters, forming a total of 20 chapters covering both translators.

The third question was about class balance. To address such question we repeated the first experiment but considered randomizing the choice of the classes for training and testing purposes, and compared the results to the balanced experiment. Since we have 30 samples which introduce 60(30x2) instances, and we are considering an imbalance problem, we choose randomly 40 out of the 60 instances for training and the rest are used for testing.

C. Classifiers:

We used six different classifiers in our experiments to study which of them can produce more accurate results in our problem. These classifiers are:

1. **FT**: Functional Tree (FT) is a classifier that was introduced by João in 2004 [40]. This classifier uses decision tests based on a combination of attributes. FT is used with its default parameters using WEKA where the minimum number of instances at which a node is considered for splitting is set to 15, and the number of fixed LogitBoost iterations is set to 15 with no weight trimming.
2. **NBTree**: This decision tree uses Naive Bayes classifiers at the leaves. This hyperid classifier is firstly presented by Kohavi in 1996 [41].
3. **Random Forest**: This classifier generates a random number of decision trees that represent a forest to be used for classification[42].

TABLE 1 EXPERIMENTS PARAMETERS

$f(x_i)$	is the frequency of each motif in a sample i for the first translator
$\Sigma(f(x_i))$	is the summation of all frequencies of motifs in a sample i for the first translator
$F(y_i)$	is the frequency of each motif in a sample i for the second translator
$\Sigma(f(y_i))$	is the summation of all frequencies of motifs in a sample i for the second translator
$F(z_i)$	is the frequency of each motif in a sample i for the original Arabic text

We used this classifier with no constraints on the maximum depth of the trees and the number of attributes to be used, while the number of trees is limited to 10.

4. **Random Tree:** While constructing decision trees, a random number of attributes is chosen at each node to introduce a Random Tree classifier. We used the defaults parameters for this classifier which includes no backfitting or pruning.
5. **FURIA:** Fuzzy Unordered Rule Induction algorithm (FURIA) is discussed in detail in section II(C) of this paper; the parameters for applying this algorithm includes setting the minimum total weight of the instances in a rule to two, with two runs of an optimization procedure, and to use check for error rate that ≥ 0.5 for the stopping criterion.
6. **FLR:** Fuzzy Lattice Reasoning Classifier (FLR) is discussed earlier in section II(C) as well. The vigilance parameter value is set to 0.5 while running our experiments.
7. **Multiclass Classifier:** This classifier uses multi-class datasets with 2-class classifiers. In these experiments we used a logistic classifier. It applies error correcting output codes to increase the accuracy.

VI. RESULTS AND ANALYSIS

Looking at Fig. 2; when we calculated the means of the 30 sample that we have for each translator, we found that the appearance of the motifs is different for each author. The second translator has the highest average for all motifs.

A. Paired T-test:

We applied paired t-test on the frequencies of motifs for each paired sample text, where the pair here represents two translations of the same original text. Two tail t-critical for $\alpha=0.05$ and sample size of 30 is 2.04523.

Table 2 displays the t-calculated for each motif. All of them show that the differences between the frequencies for the two translators are significant.

B. Correlation between motifs:

We calculated the Pearson correlation coefficient between the motifs for each translator to find how often these motifs appeared concurrently, and then compared them to each other. We summarized the important differences in Table 3.

Since these motifs represent links between words in the text, and the correlation represents how often these motifs occur together, the difference between these correlations indicates difference in translator's style.

C. Experiment 1:

For Experiment 1; as Table 4 shows the best results are obtained using the functional tree classifier and multiclass classifier with test 5 which includes the number of nodes and edges with the motif frequency as inputs. The accuracy of these classifiers is 70%, which is the same accuracy of NBTree and FLR when applied to test 3, where the data is normalised against the original text bias.

D. Experiment 2:

For experiment 2, we test using smaller training sample size, which are 10 for each translator. We noticed that the results for the multiclass classifier are enhanced compared to the first experiment, while the functional tree accuracy decreased. The best accuracy obtained for this experiment remained at 70% as shown in

Table 5. This level of accuracy is obtained using Random Tree classifier and FLR classifier for test2 where the data is normalized against the translator bias. It is also obtained by the multiclass classifier with test5 which is the same result for this classifier in the first experiment.

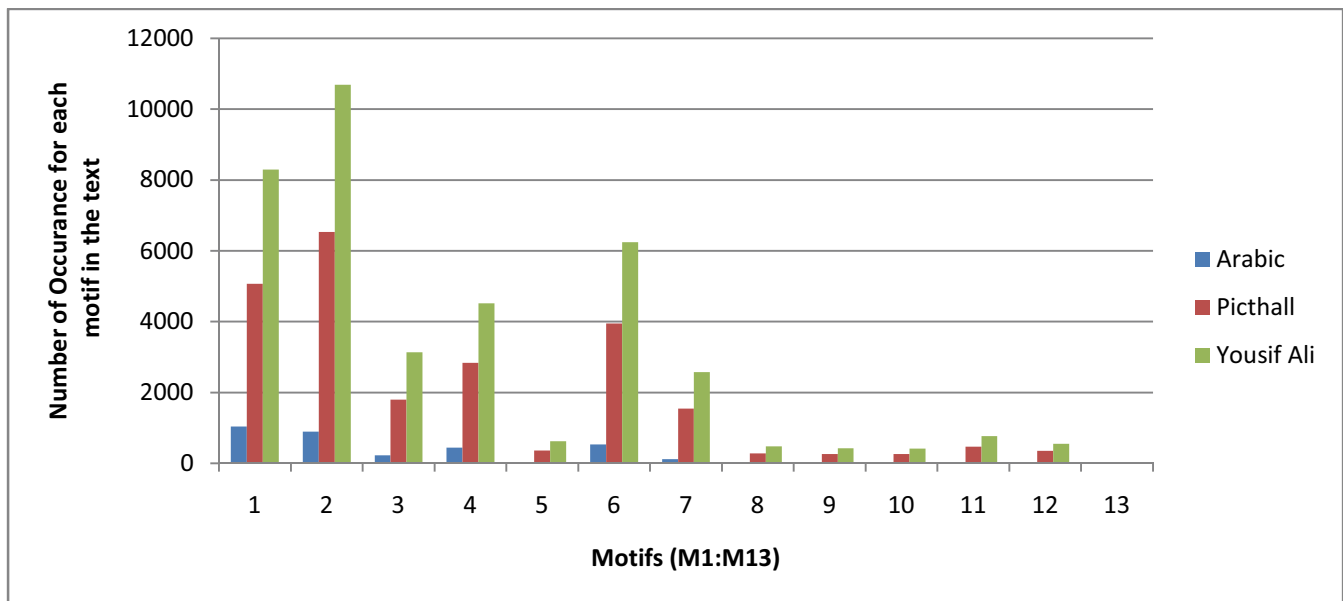


Figure 2. Comparison between the average of the 13 motifs for the Arabic, First English Translation, Second Translation

E. Experiment 3:

In the third experiment, we addressed the imbalance class problem. We used the 40 instances for training as 24 instances for the first translator “Pitchall” and 16 instances for the second translator “Yousif Ali”. For testing, we used 20 instances, where 6 instances for “Pitchall” and 14 for “Yousif Ali”. We noticed that most of the results were only 30% as shown in

Table 6 where the classifiers classified all the instances to belong to the first translator. The only classifier that gives acceptable results was FLR, where we got 70% accuracy for test 3 where data is normalized against the original text.

In general, the Random Forest classifier and FURIA consistently failed to detect the differences between translator’s styles. On the other hand, the FLR classifier introduced acceptable results in the three experiments. Test 1 and test 4 failed to identify the translator’s style, where in test 1 the data is not normalized, and in test 4 it is normalized against both the translator bias and the original text bias.

Both test 3 and test 5 introduced acceptable classifiers three times in the overall experiments. That indicates the importance of normalizing the data against its original text bias and the importance of including the number of nodes and edges into the attributes as it may help the classifier to normalize the data implicitly.

TABLE 2 PAIRED T-TEST BETWEEN FREQUENCIES OF MOTIFS FOR THE TWO TRANSLATORS

Motif	t-test
M1	-5.21757
M2	-5.36426
M3	-4.71189
M4	-5.48586
M5	-4.39642
M6	-5.46861
M7	-5.28081
M8	-4.65985
M9	-4.42133
M10	-3.93762
M11	-4.36377
M12	-4.94248
M13	-2.75729

TABLE 3 CORRELATION BETWEEN FREQUENCIES OF MOTIFS

Motifs	First Translator “Pitchall”	Second Translator “Yousif Ali”
(M2, M13)	0.768	0.523
(M4, M13)	0.910	0.550
(M6, M13)	0.816	0.495
(M7, M13)	0.911	0.640
(M9, M13)	0.946	0.711
(M11, M13)	0.978	0.751
(M12, M13)	0.975	0.752

TABLE 4 ACCURACY OF THE DIFFERENT CLASSIFIERS FOR EXPIREMENT1

Classifier	Functional Tree	NBTree	Random Tree	Random Forest	FURIA	FLR	Multiclass Classifier
Test1	60%	50%	55%	25%	50%	50%	55%
Test2	65%	50%	65%	65%	50%	60%	45%
Test3	65%	70%	45%	60%	55%	70%	60%
Test4	50%	50%	50%	50%	55%	60%	45%
Test5	70%	50%	60%	50%	50%	50%	70%

TABLE 5 ACCURACY OF THE DIFFERENT CLASSIFIERS FOR EXPIREMENT2

Classifier	Functional Tree	NBTree	Random Tree	Random Forest	FURIA	FLR	Multiclass Classifier
Test1	50%	50%	60%	55%	50%	50%	65%
Test2	50%	50%	70%	60%	60%	70%	60%
Test3	50%	50%	60%	65%	55%	65%	60%
Test4	50%	50%	60%	45%	50%	55%	60%
Test5	50%	50%	50%	65%	60%	50%	70%

TABLE 6 ACCURACY OF THE DIFFERENT CLASSIFIERS FOR EXPIREMENT3

Classifier	Functional Tree	NBTree	Random Tree	Random Forest	FURIA	FLR	Multiclass Classifier
Test1	45%	30%	25%	60%	30%	30%	40%
Test2	60%	30%	30%	20%	30%	30%	50%
Test3	45%	45%	45%	50%	55%	70%	35%
Test4	30%	30%	40%	30%	45%	40%	40%
Test5	45%	30%	50%	35%	30%	30%	45%

VII. CONCLUSION

Studying “Translator Stylometry” is a non-trivial task. While there has been research in linguistics and social sciences discussing the characteristics of each translator, this line of research is limited in computational linguistics. Contrary to previous findings that this problem does not have an automatic solution, this paper presented a first attempt to counteract this belief. We demonstrated that translators cannot disappear under the skin of the original authors; they have their own identities. Different translators represent the same idea in different ways. Although some existing authorship attributions couldn’t capture these differences [15], this work shows that we can use social network analysis to differentiate between translators’ styles. Detecting network motifs shows that each author is using certain patterns while writing. The proposed method introduces a classifier that can classify translated texts into their translators with accuracy of 70%. Fuzzy Lattice Reasoning Classifier (FLR) introduces the best results among seven tested classifiers. Normalising the data against its original text and the network size offers promising results. Further analysis is needed to explore this research area.

VIII. REFERENCES

- [1] M. Koppel, J. Schler, and S. Argamon, *Computational methods in authorship attribution*. Journal of the American Society for Information Science and Technology, 2009. **60**(1): p. 9-26.
- [2] E. Stamatatos, *A survey of modern authorship attribution methods*. Journal of the American Society for Information Science and Technology, 2009. **60**(3): p. 538-556.
- [3] S. Argamon, *Measuring the usefulness of function words for authorship attribution*. in *In Proceedings of the 2005 ACH/ALLC Conference*. 2005.
- [4] P. Fuchun, S. Dale, W. Shaojun, and K. Vlado, *Language independent authorship attribution using character level language models*, in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*. 2003, Association for Computational Linguistics: Budapest, Hungary.
- [5] A.M. García and J.C. Martín, *Function words in authorship attribution studies*. Literary Linguist Computing, 2007. **22**(1): p. 49-66.
- [6] J. Grieve, *Quantitative authorship attribution: An evaluation of techniques*. Literary and Linguistic Computing, 2007. **22**(3): p. 251-270.
- [7] H.V. Halteren, *Author verification by linguistic profiling: An exploration of the parameter space*. ACM Transactions on Speech and Language Processing (TSLP) 2007. **4**(1): p. 1-17.
- [8] P. Juola, *Authorship attribution*. Foundations and Trends in Information Retrieval, 2006. **1**(3): p. 233-334.
- [9] P. Juola, *Future trends in authorship attribution*, in *Advances in digital forensics iii*, P. Craiger and S. Sheno, Editors. 2007, Springer Boston. p. 119-132.
- [10] P. Juola, J. Sofko, and P. Brennan, *A prototype for authorship attribution studies*. Literary and Linguistic Computing, 2006. **21**(2): p. 169-178.
- [11] D. Labbe', *Experiments on authorship attribution by intertextual distance in english*. Journal of Quantitative Linguistics, 2007. **14**(1): p. 33 - 80.
- [12] D. Madigan, et al. *Author identification on the large scale*. in *In Proc. of the Meeting of the Classification Society of North America*. 2005.
- [13] J. Rudman and B. Keith, *Authorship attribution: Statistical and computational methods*, in *Encyclopedia of language & linguistics*. 2006, Elsevier: Oxford. p. 611-617.
- [14] S. Urszula and A.C. Krzysztof, *Application of artificial neural networks to stylometric analysis*, in *Proceedings of the 8th conference on Systems theory and scientific computation*. 2008, World Scientific and Engineering Academy and Society (WSEAS): Rhodes, Greece.
- [15] M. Mihailov and M. Villikka, *Is there such a thing as a translator's style?* . in *Proceedings of the Corpus Linguistics*. 2001. Lancaster.
- [16] R. Milo, et al., *Network motifs: Simple building blocks of complex networks*. Science, 2002. **298**(5594): p. 824-827.
- [17] M. Winters, *From modal particles to point of view a theoretical framework for the analysis of translator attitude*. Translation and Interpreting Studies, 2010. **5**: p. 163-185.
- [18] X. Xiumei, *Style is the relationship a relevance-theoretic approach to the translator's style*. Babel, 2006. **52**: p. 334-348.
- [19] P.S. Angermeyer, *Translation style and participant roles in court interpreting*. Journal of Sociolinguistics, 2009. **13**(1): p. 3-28.
- [20] M. Newman, *Networks: An introduction*. 2010, New York: Oxford University Press.
- [21] S. Wasserman and K. Faust, *Social network analysis: Methods and applications (structural analysis in the social sciences)*. 1994: Cambridge University Press.
- [22] U. Alon, *Network motifs: Theory and experimental approaches*. Nat Rev Genet, 2007. **8**(6): p. 450-461.
- [23] O. Sporns and R. Kötter, *Motifs in brain networks*. PLoS Biol, 2004. **2**(11): p. e369.
- [24] N. Puig, I. Pournara, and L. Wernisch, *Statistical model comparison applied to common network motifs*. BMC Systems Biology, 2010. **4**(1): p. 18.
- [25] A. Ghoneim, H. Abbass, and M. Barlow, *Characterizing game dynamics in two-player strategy games using network motifs*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2008. **38**(3): p. 682-690.
- [26] S. Itzkovitz, et al., *Coarse-graining and self-dissimilarity of complex networks*. Phys Rev E Stat Nonlin Soft Matter Phys, 2005. **71**(1 Pt 2).
- [27] S. Valverde and R.V. Solé, *Network motifs in computational graphs: A case study in software architecture*. Physical Review E, 2005. **72**(2): p. 026107.
- [28] J. Hühn and E. Hüllermeier, *Furia: An algorithm for unordered fuzzy rule induction*. Data Mining and Knowledge Discovery, 2009. **19**(3): p. 293-319.
- [29] W.C. William, *Fast effective rule induction*, in *Proceedings of the 12th international conference on machine learning (ICML)*. 1995: Tahoe City. p. 115-123.
- [30] J. Hühn and E. Hüllermeier, *An analysis of the FURIA algorithm for fuzzy rule induction*, in *Advances in machine learning i*, J. Koronacki, et al., Editors. 2010, Springer Berlin / Heidelberg. p. 321-344.
- [31] I.N. Athanasiadis, V.G. Kaburlasos, P.A. Mitkas, and V. Petridis. *Applying machine learning techniques on air quality data for real-time decision support*. in *First International NAISO Symposium on Information Technologies in Environmental Engineering (ITEE'2003)*. 2003. Gdansk, Poland: ICSC-NAISO Publishers.
- [32] I. Jenhani, N.B. Amor, and Z. Elouedi, *Decision trees as possibilistic classifiers*. International Journal of Approximate Reasoning, 2008. **48**(3): p. 784-807.
- [33] S. Procházka, *Arabic*, in *Encyclopedia of language & linguistics*, B. Keith, Editor. 2006, Elsevier: Oxford. p. 423-431.
- [34] M. Abdel Haleem, *Islam and Arabic*, in *Encyclopedia of language & linguistics*, B. Keith, Editor. 2006, Elsevier: Oxford. p. 34-37.
- [35] K. Mohammed, *Assessing English translations of the Qur'an*. The Middle East Quarterly, 2005. **XII**(2): p. 58-71.
- [36] A. Abbasi and H. Chen, *Applying authorship analysis to extremist-group web forum messages*. 2005. **20**(5): p. 67-75.
- [37] F. Schreiber and H. Schwöbbermeyer, *MAVisto: A tool for the exploration of network motifs*. Bioinformatics, 2005. **21**(17): p. 3572-3574.
- [38] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks*. Science, 2002. **296**(5569): p. 910-913.
- [39] S. Maslov, K. Sneppen, and U. Alon, *Correlation profiles and motifs in complex networks*. Handbook of graphs and networks. 2005: Wiley-VCH Verlag GmbH & Co. KGaA. 168-198.
- [40] G. João, *Functional trees*. Machine Learning, 2004. **55**(3): p. 219-250.
- [41] R. Kohavi. *Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid*. 1996.
- [42] L. Breiman, *Random forests*. Mach. Learn., 2001. **45**(1): p. 5-32.