IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2021, accepted June 24, 2021, date of publication June 29, 2021, date of current version July 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093370

On Explainable Features for Translatorship Attribution: Unveiling the Translator's Style With Causality

CHRISTIAN CABALLERO, HIRAM CALVO, AND ILDAR BATYRSHIN, (Senior Member, IEEE)
Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City 07738, Mexico

Corresponding author: Hiram Calvo (hcalvo@cic.ipn.mx)

This work was supported in part by the Mexican Government through Consejo Nacional de Ciencia y Tecnología (CONACYT) under CONACYT-SNI and CONACYT-Becas Nacionales; and in part by the Instituto Politécnico Nacional (IPN) under Multidisciplinary Project 2083, Project SIP 20210189, Project SIP 20211874, COFAA-SIBE, and BEIFI-IPN.

This work did not involve human subjects or animals in its research.

ABSTRACT Translatorship attribution deals with accurately attributing a translation to its translator. The task is challenging because several factors can confound the attribution such as the original author's style, genre, and topic of the text. The attribution and the identification of the translator's style could contribute to fields including translation studies and forensic linguistics. In this paper, we pose translatorship attribution as a multiclass classification problem and employ machine learning algorithms. To address the problem of confounding, we use corpora of English translations of the same source material (parallel corpora) to identify the translators' personal style. We propose two novel feature sets in this task: i) a list of cohesive markers with and without their surrounding punctuation and ii) syntactic *n*-grams to capture real syntactic information. We employ χ^2 feature selection and, using 10-fold cross-validation, assess the accuracy of several classifiers trained with our proposed features and with word, punctuation, POS, and POS-punctuation *n*-grams. The results show that the proposed features yield comparable and even higher accuracy results than the reported in the literature on the same corpora and prove that POS-punctuation *n*-grams are an effective feature set for this task. We also recover the most distinctive features and provide examples of stylistic interpretations of them for each translator. Finally, using insights from causal inference, where confounding is well-defined and studied, we provide a novel explanation for the accepted need of using parallel and contemporaneous corpora on this task and for the different results among types of features.

INDEX TERMS Computational linguistics, translator style, stylometry, machine learning, causal inference.

I. INTRODUCTION

Attributing a piece of text to a certain author (i.e., authorship attribution) is a well-established task in computational linguistics (for example, see [1]–[4]). However, attributing a translated document to its *translator* (we might call it translatorship attribution) is a task studied and reported only in a few papers and with only a subset of them reporting positive results.

Results for this problem can contribute to, for example, the fields of translation studies, education, intellectual property, and forensic linguistics. In these fields, it might be important to solve controversies regarding the original translatorship of a given text or to determine if a translation was plagiarized.

Translatorship attribution can be traced back to the seminal paper of Baker [5]. In it, she proposed an outline of a framework to study the style of a translator or a group of translators. ''Style'', as she referred to it, ''is a matter of patterning: it involves describing preferred or recurring patterns of linguistic behavior, rather than individual or one-off instances of intervention'' [5, p. 245].

This definition of style would include both conscious and unconscious *choices* made by the writer and, as in the current case, the translator. Thus, the style of a translator has more ''room'' to emerge when there are different options to choose from. Some of the unconscious elements of translatorial style are, for example, the influence of the target and

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

source languages, the education and idiolect of the translator, and the use of high-frequency function words as outlined in [6, p. 180].

We can see the task as twofold: on the one hand, accurately perform the attribution, and, on the other hand, recover the unconscious features (i.e., the style) useful for the correct attribution. Baker [5] suggested exploring whether the translator's style might be manifested in the preferences of using ''specific lexical items, syntactic patterns, cohesive devices, or even style of punctuation'' [5, p. 248]. Despite this suggestion, up to this day the published research addressing the problem has used features typically used in automatic authorship attribution with mixed results and none has studied the influence of different kinds of features on the attribution.

In this paper, we see the translatorship attribution task as a multiclass classification problem. Following Baker's suggestions, we propose to use cohesive markers and punctuation as well as syntactic information as features to be used along with machine learning techniques.

Cohesive markers work at the linguistic level of *discourse*. Discourse is the ''study of units of language and language use consisting of more than a single sentence, but connected by some system of related topics'' [7, p. 388]. In general, discourse markers relate to textual cohesion, working as textual linking devices [8, p. 17]. *Cohesion* and *coherence* are two often confused and closely related discourse phenomena. Cohesion is a property of text, whereas coherence pertains to discourse and is derived within the process of instantiation of the interpretation potential of a text. In other words, cohesion fosters coherence by means of cohesive devices that guide the reader in the processing of text [9, p. 11].

Cohesive markers satisfy functions within the text such as elaboration, contrasting, and summarization. Since different natural languages have a different set of cohesive markers, they are a good place for the free choice of the translator to emerge. We gathered a list of English cohesive markers and used a bag-of-items model of them to represent each document in a vector space model. In other words, we counted the number of occurrences of each cohesive marker in each document to model the document as a vector. We repeated this procedure with the same list of cohesive markers but this time including the surrounding punctuation marks to have a representation that uses these discourse markers plus the punctuation the translator chose to introduce them in the text.

In order to consider actual syntactic information, we used syntactic n-grams (as described in [10]) to represent the documents using a bag-of-items model. Syntactic n-grams are n adjacent elements within a sentence *traversing* the dependency tree. Thus, the documents must be parsed first. Parsing tends to be computation-intensive. However, modern tools allow performing this analysis in a personal computer with no special hardware requirements very fast. To provide an example, the parsing of our over one million words corpora (see Subsection III-A) took less than 15 minutes in a modern laptop.

We also represented the documents with word, POS, POS-punctuation, and only punctuation n-grams so we could compare the performance of our proposed features. We carried out classification experiments for each kind of feature set using four ''classic'' machine learning algorithms and used accuracy as a metric (since the corpora are balanced among classes) to assess the performance of each feature set and each classification algorithm combination. The election on the classifiers as well as of the corpora used was based on the state of the art to ease comparison (see Subsection II-B). The goal was to assess whether our proposed features could compare with the ones used in the literature and not so much on beating the performance accuracy, which is already high on the state of the art for the corpora we used.

Additionally, as we had already mentioned, we see translatorship attribution as twofold: attribution and personal translator's style. Therefore, the use of classic machine learning classifiers (such as Logistic Regression) allows us to recover the features that proved most distinctive for each translator. We provide a stylistic interpretation of these features (see Section V). Therefore we excluded the use of more modern machine learning techniques (such as Deep Learning) and the use of feature extraction techniques (such as Principal Component Analysis) that lose the human-interpretable quality of the features [11, Ch. 9].

Given the nature of the task, many variables can bias the results. For this reason, it has been suggested to use parallel and contemporaneous corpora (see [5], [6], [12]) to disentangle the translator's style from these other *confounding* variables such as the original author's influence, source and target languages, textual genre, topic, and period of the translation.

Causal inference provides a formal and graphical framework for studying confounding (see [13, Ch. 7]). Using insights from causal inference, we propose a causal diagram for explaining the apparent need of using parallel and contemporaneous corpora in translatorship attribution and provide a justification for using some features even when the corpus is not parallel or contemporaneous.

Summarizing, in this paper we propose a novel set of features to use in the understudied task of translatorship attribution: cohesive markers along with punctuation. Additionally, we use syntactic n-grams (as proposed in [10]) as features in order to capture syntactic information to perform the translator attribution for the first time in this task. Furthermore, we recover the most relevant features for the attribution and provide a stylistic interpretation for different sets of features. Lastly, we propose a causal explanation for the difference in performance for different kinds of features and the nature of the corpus recommended for the task (i.e., parallel and contemporaneous).

The rest of the paper is structured as follows: in Section II, we review related research in translator stylometry; in Section III, we describe the corpora used to carry out our experiments, the preprocessing, processing, and classification steps as well as the methodology followed to recover the most distinctive features for each translator—their stylistic

''fingerprints''; in Section IV, we present the results of our experiments and compare them with the results in the state of the art, which used the same corpora; in Section V, we discuss the results for each corpus and present a causal explanation for using parallel and contemporaneous corpora in this task as well as the differences in performance we found for some features; lastly, in Section VI, we give our conclusions and plans for future work.

II. RELATED WORK

In this section, we will first describe earlier research done on translated text, followed by the most recent works on translator stylometry on which we based our research.

A. PREVIOUS WORK ON TRANSLATED TEXTS

As we mentioned in Section I, Baker [5] drew attention to the problem of going after an individual translator's style. It is worth mentioning that, as part of her research, she and her team built a corpus containing around ten million words of English text translated from several source languages, called the Translational English Corpus (TEC).[1] From this corpus, she extracted English translations of two translators for her paper: six texts (one from Portuguese and the remaining five from Spanish) of one translator and three texts from Arabic of the second translator.

With those texts, she compared the variation in type/token ratio (using a moving average of 1000 words), average sentence length, and reporting structures (realizations of the reporting verb *say*) finding differences between the translators and consistencies across their individual production. Later, she proceeded to give plausible explanations for the differences and concluded hinting possible directions to develop further the methodology for such a study in order to try to *fix* or control for more variables by ''comparing different translations of the same source text into the same target language, by different translators, thus keeping the variables of author and source language constant'' [5].

Other work around the same time as Baker's paper posed the question of whether translators have ''stylistic fingerprints'' [14]. In this regard, results were not conclusive since the translations were more similar to the original works than among different translations by the same translator.

More recently, other works dealt with translated texts (see [15]–[18]) finding that the translators leave behind traces on the texts. In particular, [18] provided an explanation for the discrepancies found in earlier work trying to find the translator's style. Lee [18] proposed that the structural distance of the pair of working languages influences the room the translator has to showcase their creativity and tested his hypothesis using English translations of French and Korean novels.

Lastly, Covington *et al.* [19] analyzed several biblical passages translated into present-day English using as features mean sentence length, vocabulary diversity (as the moving average of type/token ratio), and idea density (number of propositions divided by the number of words) in order to perform a clustering analysis. Their results show that the translations cluster by the intention of the translation (e.g., preserve wording of King James Version, smooth reading, and elegant literary style). These clusters mirror the literary history of the translations and show another application of stylometric analysis of translations.

B. STATE OF THE ART ON TRANSLATOR STYLOMETRY

The two most recent works, which are also closer to what we attempt here, are [20] and [21]. The authors of both papers use the methodology suggested by Baker [5] of using parallel corpora—same source text translated by different people into the same target language. Also, both teams employ machine learning algorithms and deviate from the use of most frequent words and/or Burrows' Delta and its modifications such as Eder's Delta [22] as their primary approach as it had been prevalent in the works discussed above in Subsection II-A.

Lynch and Vogel [20] decided to use two corpora. The first one comprises seven works by Norwegian playwright Henrik Ibsen translated into English by two translators. One play, *Ghosts*, was translated by both translators, and the researchers used it as a parallel corpus in their first experiment, which consisted in training several Support Vector Machine (SVM) classifiers with the ten most distinctive words, ten most distinctive word bigrams, and ten most distinctive Part-of-Speech (POS) bigrams. The most distinctive features of each type were obtained using the χ^2 statistic. Later they tested this method on the remaining six plays (a nonparallel corpus) and lastly they trained other machine learning classifiers (Naïve Bayes, Decision Tree, and Logistic Regression) using 18 document level features such as average sentence length, type/token ratio, average word length, among others.

The second corpus is composed of ten works from Russian writer Anton Chekhov translated into English by two female translators. Out of the ten, six were translated by both. They repeated their methodology on this corpus. By the end, they show a clustering analysis using a modification of Burrows' Delta [22] with the 100 most frequent words and with the ten most distinctive words on both corpora. Only using the ten most distinctive words they achieve a perfect clustering by translator instead of by work, which confirms the usefulness of said set of distinctive words in the classification task.

In the second work, El-Fiqi *et al.* [21] also used two corpora: the last six (out of 30) parts of seven translations into English from Arabic of the Quran and three translations into English from Spanish of the two parts of the novel *Don Quixote* by Miguel de Cervantes. After obtaining negative results using the methods in [14], they use *network motifs* (repeated subgraphs on the word adjacency network), which, they argue, capture some syntactic patterns in the writing of each translator.

This latter approach gives them a higher accuracy in the classification task after they performed feature selection

[1] https://genealogiesofknowledge.net/translational-english-corpus-tec/

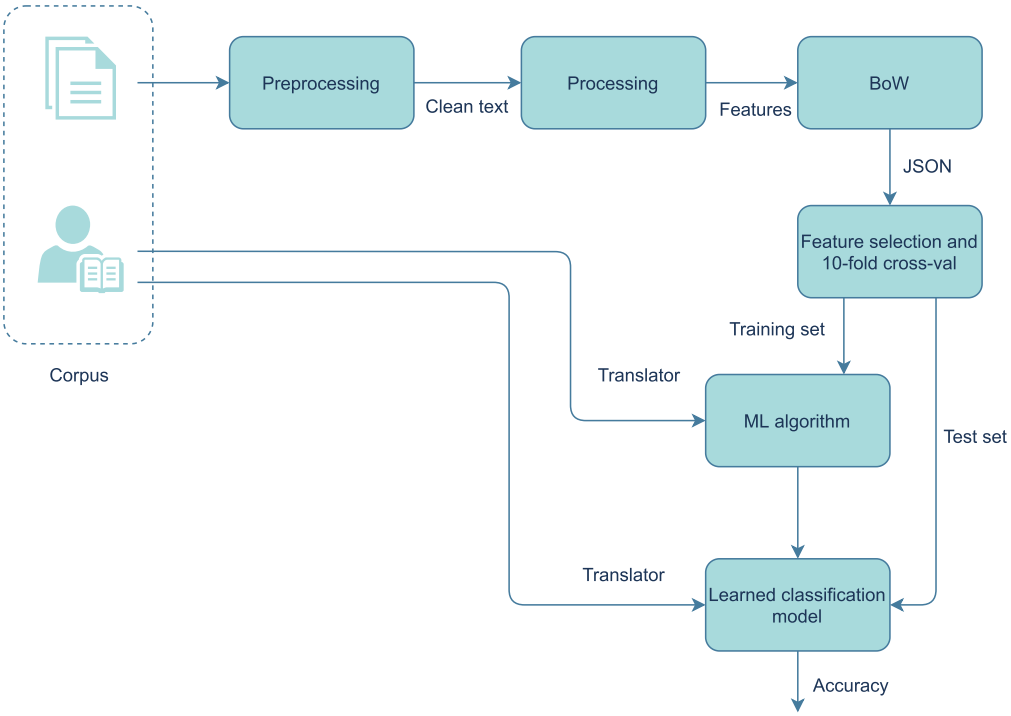

FIGURE 1. Workflow. In order to obtain the accuracy of each feature set, the translations were preprocessed, processed, modeled as bag-of-words (BoW) to then perform a 10-fold cross-validation with χ^2 feature selection.

by ranking. It is worth mentioning in order to contrast their results against ours, that, as part of their data preprocessing, they cleaned and lemmatized the text leaving only alphanumeric characters (thus removing punctuation). Contrasting with [20] and even though they did use a feature selection technique by ranking, El-Fiqi *et al.* [21] did not present the network motifs most distinctive for each translator nor provide a stylistic interpretation of their results.

III. METHODOLOGY

This section details the corpora and the steps carried out to perform the experiments (see Fig. 1) to assess the accuracy of the attribution as well as the procedure to recover the most distinctive features for each translator. We used the Python programming language[2] in all the steps of the process. Additionally, we employed spaCy, a library for natural language processing (NLP) [23]; scikit-learn, a library for machine learning [24], and pandas, a library for data analysis and handling [25], [26].

A. CORPORA

We experimented with one corpus from each of the two works reviewed in Subsection II-B: the Ibsen corpus from [20] and the *Don Quixote* corpus from [21]. We downloaded the Ibsen plays (translations of theater plays from Norwegian into English) from the Project Gutenberg[3] website. This corpus is detailed in Table 1. Regarding the three translations of the two parts of *Don Quixote* (three translations of the famous novel from Spanish into English), we retrieved the texts from Professor Hussein Abbass' website.[4] This corpus is summarized in Table 2 and was already segmented by chapter and by translator.

On average, the Ibsen corpus has 130,000 words per translator whereas the *Don Quixote* corpus has around 405,000 words per translator. In addition, note that in the Ibsen corpus, only one play, *Ghosts*, was translated by both translators meanwhile the other six plays were translated

TABLE 1. Number of words per play in the Ibsen corpus.

Translator	Play	Total words
William Archer	*Ghosts*	24,266
	John Gabriel Borkman	26,038
	Little Eyolf	21,646
	When We Dead Awaken	19,055
R. Farquharson Sharp	*An Enemy of the People*	31,173
	Ghosts	22,510
	Pillars of Society	31,984
	Rosmersholm	27,394

TABLE 2. Number of words per chapter in the *Don Quixote* corpus.

Translator	Part I 53 chapters	Part II 74 chapters
Charles Jarvis	1,700-8,202	797-5,530
John Ormsby	1,648-8,304	823-5,680
Thomas Shelton	1,820-8,921	759-5,158

[2]https://www.python.org
[3]https://www.gutenberg.org
[4]http://www.husseinabbass.net/translator.html

three by one translator and the other three by the other. We treated these texts as two subcorpora: one parallel (i.e., the play *Ghosts*) and one nonparallel (i.e., the remaining six plays in Table 1).

B. PREPROCESSING

For the two corpora mentioned in detail in Section III-A, we replaced the special characters with their plain counterparts (e.g., *ë* for *e*); removed numbers between brackets, used for footnotes, and replaced all white space, including carriage returns with only one simple white space. In particular, for the Ibsen corpus, we removed Project Gutenberg's front matter and legal information and segmented the plays into 5 kB chunks as in [20] in order to have more (albeit smaller) samples for each play. In addition, all square brackets were replaced with parentheses since one translator used exclusively parentheses whereas the other used square brackets for stage directions rendering the classification trivial when using punctuation.

C. PROCESSING

With the text files already clean and segmented (per chapter for the *Don Quixote* corpus and in 5 kB segments for the Ibsen corpus), the translations were processed using the specialized NLP Python library spaCy for tokenization, POS-tagging, and dependency parsing. This library comes with available Convolutional Neural Network (CNN) language models trained on different sources for POS-tagging, dependency parsing, and named-entity recognition (NER) for a variety of natural languages.

D. FEATURES

We used different feature sets to represent the documents in a vector space model using a simple bag-of-items representation with raw counts for each feature in each file:

- Word *n*-grams for $n \in \{1, 2, 3\}$ transforming every character to lowercase and replacing all proper nouns with the POS-tag "PROPN" to avoid the identification of the translator by idiosyncratic choices on characters' names
- POS *n*-grams for $n \in \{2, 3\}$ ignoring punctuation marks
- Punctuation *n*-grams masking all words with the character "*"
- POS-punctuation *n*-grams for $n \in \{2, 3\}$
- Cohesive markers—mainly taken from Professor David O'Regan's website[5]—with and without the surrounding punctuation marks
- Syntactic word *n*-grams for $n \in \{2, 3\}$ using the metalanguage described in [10] and once again masking all proper nouns with their POS-tag "PROPN"

In Table 3, we can appreciate the intuition behind using cohesive markers and punctuation marks (our novel proposal). The differences in choices of cohesive markers and punctuation between two translations of the same excerpt of *Don Quixote* are in boldface. The cohesive markers satisfy a function at the discourse level, such as exemplification, summarization, contrast, etc., and the inventory varies among natural languages as well, giving freedom to the translator to choose a particular marker at the discourse level and not at the lexical one.

TABLE 3. Two translations of the same excerpt from Chapter 12 of the first part of *Don Quixote*.

Jarvis	Ormsby
Don Quixote desired Pedro to tell him who the deceased was**, and** who that shepherdess**. To which** Pedro answered**, that** all he knew was**, that** the deceased was a wealthy gentleman of a neighbouring village among the hills thereabout**, who** had studied many years in Salamanca**; at the end of which** time he returned home with the character of a very knowing and well-read person: **particularly** it was said, he understood the science of the stars**, and** what the sun and moon are doing in the sky: for he told us punctually the clipse of the sun and moon."	Don Quixote asked Pedro to tell him who the dead man was **and** who the shepherdess**, to which** Pedro replied **that** all he knew was **that** the dead man was a wealthy gentleman belonging to a village in those mountains**, who** had been a student at Salamanca for many years**, at the end of which** he returned to his village with the reputation of being very learned and deeply read**. "Above all**, they said, he was learned in the science of the stars **and** of what went on yonder in the heavens and the sun and the moon, for he told us of the cris of the sun and moon to exact time."

The words in boldface are contrasting uses of some cohesive markers and their surrounding punctuation marks.

The bag-of-words representation for each text file along with the corresponding label for its translator for all files in each corpus was saved to disk to JSON files—one JSON file per corpus and kind of features. We used these JSON files to train the machine learning classifiers.

E. CLASSIFIERS

We chose four machine learning classifiers: Support Vector Machine Classifier (SVC) with a linear kernel, Naïve Bayes (NB), Decision Tree (DT), and Logistic Regression (LR). Additionally, we employed *k*-fold cross-validation as a validation technique to compute the accuracy of each classifier. The choice of using these four classifiers is driven by two criteria.

The first is that these are the algorithms used in the state of the art (see Subsection II-B) and therefore it is easier to assess the impact of using different feature sets if we are using the same corpora and the same algorithms.

The second criterion is that we are particularly interested in recovering the most distinctive features for the attribution to each translator to provide a stylistic interpretation. We explain below in the next subsection how we recovered the most distinctive features for each translator.

Given these two criteria, we did not consider using more modern machine learning techniques, such as Deep Learning or ensemble methods since they are not so easy to interpret [27].

We used the classifiers' default values in scikit-learn. One default setting important to mention is the strategy for training

[5]http://home.ku.edu.tr/~doregan/Writing/Cohesion.html Last accessed on March 15, 2021

the classifiers when we have a multiclass classification problem (also called multinomial [28, Ch. 8]) and a binary classifier; the version of scikit-learn we used (0.21) has a one-vs-all strategy for default in which a classifier is trained per each class (see [29, Ch. 3] for a detailed description).

As part of the pipeline of the SVC and the LR, we standardized the values of the features to have a standard deviation of 1. We did not subtract the mean in order to not break the sparse representation of the data, which is an efficient way scikit-learn uses to represent sparse matrices (matrices with many zeros) in memory.

Lastly, we need to mention that the inherent high dimensionality is one drawback of using the bag-of-items representation we mentioned in Subsection III-D. We added to the pipeline of each model a dimensionality reduction technique to address this issue. There are several dimensionality reduction techniques: some transform the original features (e.g., Principal Component Analysis or Linear Discriminant Analysis) and others keep the form of the original features, but choose the most useful (e.g., Variance Thresholding or χ^2) [11, Ch. 9 and 10]. To fulfill the second criterion we just mentioned above, we employed a feature selection technique.

Feature selection is still an open research area. For example, recent developments use bioinspired evolutionary processes [30], [31] for optimizing many objectives simultaneously (e.g., minimizing the number of features while maximizing or maintaining accuracy or information gain). Interestingly for the present work, there is even research done in feature selection for text mining using causal inference [32], where the goal is to go beyond mere association between the occurrence of a feature and a given class or label and find ''causal'' relationships (i.e., which features *cause* a document to belong to a given class).

Nevertheless, we decided to use the simple χ^2 statistic as a feature selection technique within the cross-validation. First, because that is the technique used in [20], which facilitates the direct comparison between their results and ours. Second, since the motivation of this work is to assess the performance of some feature sets in known corpora, we decided to use a time-tested technique in text mining.

This technique computes a statistic that measures the deviation of the occurrence of a feature and a label if the two were independent. By ranking the features by this statistic, we can select the k features most relevant for the classification. This technique reduces the noise from using too many features that may not contribute to the classification and reduces the chances of overfitting the data due to the reduced dimensionality [33, Ch. 13]. For all of our experiments, we selected the $k = 25$ best features.

F. RECOVERING THE MOST RELEVANT FEATURES

We recover the features from the *learned* classifiers once we know they are accurate. SVC with a linear kernel, NB, and LR allow recovering the coefficients for each feature, whereas DT allows to recover the features (Gini) importance.

The classifier is trained with all the samples after performing feature selection. What we recovered were the n features with the largest coefficients for each class. We used $n = 10$ and the LR classifier. We can see the linear combination of the features as an expression for the log-odds for each class (i.e., the translator) for this algorithm [27, Ch. 4].

IV. EXPERIMENTS AND RESULTS

In this section, we present the results of the different experiments we performed in order to assess the accuracy of the classification. We also compare the accuracy of our classifiers with the results presented in the literature that used the same corpora.

As we mentioned above in Section I, the goal of the experiments was to assess if our proposed features yield comparable results with the features used in the state of the art. To assess it, we performed individual experiments for each of the feature sets described in Subsection III-D with each of the classifiers presented in Subsection III-E in the corpora described in Subsection III-A.

The order of the experiments is the following:

1) Subsection IV-A presents the results of the experiments performed on the *Don Quixote* corpus. We contrast them with the results in [21].
2) Then, in Subsection IV-B, we present the results for the experiments on the Ibsen parallel corpus (i.e., the two translations of the play *Ghosts*) and compare them with the results on the same corpus of [20].
3) Next, in Subsection IV-C, we show the results for the Ibsen nonparallel corpus (i.e., the remaining six plays) and again we compare them with the results in [20].
4) Lastly, in Subsection IV-D, we present the results of two sets of experiments. We train on the Ibsen parallel corpus and test on the nonparallel and *vice versa*. Again, we compare our results with the results of the same kind of experiments in [20].

All the experiments were performed on a personal computer without any special hardware requirement. In particular, the most computing-intensive step is the representation of the documents using syntactic n-grams since the texts need to be parsed beforehand. The parser used (i.e., a spaCy English model) parses both corpora (well over one million words) in less than 15 minutes on personal computer. The results and the code to reproduce them are available as a GitHub repository[6] with the option of using Google Colaboratory[7] to reproduce the results on the Cloud.

A. EXPERIMENTS WITH THE DON QUIXOTE CORPUS

In Table 4, we present the mean accuracy of the 10-fold cross-validation procedure with χ^2 feature selection for the best $k = 25$ features for each of the feature sets discussed in Subsection III-D and for each of the four classifiers

[6]https://github.com/ccaballeroh/translatorship-attribution
[7]https://colab.research.google.com

TABLE 4. Mean accuracy results on the *Don Quixote* corpus using χ^2 feature selection ($k = 25$) and 10-fold cross-validation.

Features	SVC (%)	NB (%)	DT (%)	LR (%)
word unigrams	**98.15**	95.78	92.34	97.36
word bigrams	96.02	95.22	86.79	**96.56**
word trigrams	**88.08**	85.71	81.47	**88.08**
punct unigrams	96.83	**97.89**	94.96	97.09
punct bigrams	**98.42**	98.15	98.41	98.42
punct trigrams	**98.68**	97.35	97.35	97.88
POS bigrams	83.30	80.41	58.11	**84.36**
POS trigrams	82.02	81.23	60.31	**84.13**
POS-punct bigrams	96.56	95.77	94.99	**97.62**
POS-punct trigrams	96.83	**97.36**	95.79	96.57
cohesive	89.13	**90.69**	77.74	89.67
cohesive w/punct	96.00	95.75	81.75	**98.13**
syntactic bigrams	94.97	93.92	88.09	**96.04**
syntactic trigrams	85.41	84.89	74.08	**85.94**

mentioned in Subsection III-E for the *Don Quixote* corpus. The highest results per feature set are in boldface.

As mentioned in Subsection II-B, El-Fiqi *et al.* used this corpus. For this corpus, the highest mean accuracy they obtained is 77.14% using a Support Vector Machine (SVM) and their proposed features of network motifs. Later, they improved their results to 94.80% for an SVM and 95.10% for a decision tree after performing ranking of features as a method of feature selection. Their features, they argue, capture syntactic information [21, p. 27] although they do not perform a syntactic analysis.

Our method of feature selection combined with word, punctuation, and POS-punctuation *n*-grams yields comparable and, in some instances, higher results than theirs. Important for this work is the result using the proposed cohesive markers along with their surrounding punctuation, which gives a mean accuracy of 98.13% with an LR classifier, as well as the syntactic bigrams with a 96.04% accuracy and syntactic trigrams with 85.94%, which capture true syntactic information since a dependency tree is generated before building the syntactic *n*-grams.

It is worth mentioning that the results for the punctuation unigrams are misleading since our POS-tagger made some mistakes with some archaic word forms (for example, ''wing'd'', ''hath'', and ''talkest''). However, for the rest of the feature sets, the results hold.

B. EXPERIMENTS WITH THE PARALLEL IBSEN CORPUS

The results for the parallel Ibsen subcorpus (consisting of the translations of *Ghosts*) are shown in Table 5. In Subsection II-B, we mentioned that Lynch and Vogel [20] proposed and used the Ibsen corpus.

Using the ten most distinctive word unigrams and word bigrams, they reached an accuracy of 91% and 93% respectively. Using the same kind of features, our results are lower than theirs with 84% and 78% respectively. One possible explanation for the difference between their results and ours is that their tokenization is different. For example, for them, ''I've'' is one token whereas for us they are two tokens.

TABLE 5. Mean accuracy results on the Ibsen parallel corpus (i.e., *Ghosts*) using χ^2 feature selection ($k = 25$) and 10-fold cross-validation.

Features	SVC (%)	NB (%)	DT (%)	LR (%)
word unigrams	68.00	81.50	61.00	**84.00**
word bigrams	75.50	75.50	72.00	**78.00**
word trigrams	59.50	**60.00**	47.00	57.50
punct unigrams	**90.00**	81.50	74.50	85.50
punct bigrams	94.00	**98.00**	95.50	**98.00**
punct trigrams	**100.0**	**100.0**	93.50	**100.0**
POS bigrams	**62.00**	51.50	57.50	59.50
POS trigrams	24.00	22.50	**37.00**	28.00
POS-punct bigrams	**100.0**	**100.0**	93.50	97.50
POS-punct trigrams	**100.0**	**100.0**	93.50	**100.0**
cohesive	47.50	**61.50**	61.00	53.50
cohesive w/punct	**67.50**	61.50	63.50	59.50
syntactic bigrams	44.50	56.50	**83.50**	46.00
syntactic trigrams	51.50	**55.50**	53.00	53.50

TABLE 6. Mean accuracy results on the Ibsen nonparallel corpus (i.e., remaining six plays) using χ^2 feature selection ($k = 25$) and 10-fold cross-validation.

Features	SVC (%)	NB (%)	DT (%)	LR (%)
word unigrams	89.05	**91.93**	80.33	90.23
word bigrams	93.07	93.10	81.47	**94.25**
word trigrams	90.10	**90.72**	80.85	89.48
punct unigrams	91.86	79.77	86.60	**93.04**
punct bigrams	98.82	**99.41**	98.27	**99.41**
punct trigrams	98.82	**99.41**	98.27	98.82
POS bigrams	83.82	85.59	72.88	**86.21**
POS trigrams	84.44	**87.39**	67.75	83.86
POS-punct bigrams	98.82	**99.41**	98.27	**99.41**
POS-punct trigrams	**99.41**	95.95	97.06	**99.41**
cohesive	81.96	**83.20**	71.11	80.85
cohesive w/punct	84.35	84.90	83.73	**89.02**
syntactic bigrams	88.99	91.93	81.47	**92.52**
syntactic trigrams	90.82	**92.52**	86.73	92.48

They also used POS bigrams; however, they did not report their accuracy results. Using the 18 document-level features, they report 75% accuracy with an SVM and 77% for an LR classifier.

In spite of our lower results with word *n*-grams, using topic-independent features (as their 18 document-level feature set is) such as POS-punctuation *n*-grams and punctuation trigrams yields perfect accuracy. Punctuation bigrams also scored high with a 98% accuracy.

Unfortunately, our proposed sets of features (cohesive markers and syntactic *n*-grams) do not perform as well as for the previous corpus. We hypothesize that the size of the corpus and/or the genre could be behind the difference in performance.

C. EXPERIMENTS WITH THE NONPARALLEL IBSEN CORPUS

The results for the nonparallel Ibsen subcorpus (consisting of the six plays that are not *Ghosts*) are shown in Table 6.

For this subcorpus, Lynch and Vogel report accuracies of 97.5% and 95% for word unigrams and bigrams respectively. Using the same kind of features, our results are slightly

TABLE 7. Accuracy results of training with the parallel Ibsen corpus and testing on the nonparallel corpus using χ^2 feature selection ($k = 25$).

Features	SVC (%)	NB (%)	DT (%)	LR (%)
word unigrams	66.47	72.25	**78.03**	53.76
word bigrams	55.49	56.65	**60.69**	51.45
word trigrams	63.01	**68.79**	65.32	46.24
punct unigrams	39.31	43.93	**86.13**	65.32
punct bigrams	78.03	77.46	**94.22**	79.77
punct trigrams	94.22	97.11	**98.84**	76.88
POS bigrams	59.54	**67.05**	58.96	66.47
POS trigrams	53.18	53.76	53.76	**55.49**
POS-punct bigrams	94.80	93.06	**95.95**	78.03
POS-punct trigrams	**97.69**	97.11	94.80	79.19
cohesive	58.38	57.23	**64.74**	57.80
cohesive w/punct	59.54	**62.43**	61.85	61.27
syntactic bigrams	**64.74**	64.16	58.96	57.23
syntactic trigrams	**57.23**	**57.23**	53.18	49.71

lower with 91.93% and 94.25%. On the other hand, the syntactic n-grams gave a 92.52% accuracy.

Just as they point out, since the plays are different, even common nouns would be discriminating due to the plays having a different topic each. When they removed the common nouns, their results dropped to 84%. Among the top 10 features for unigrams and bigrams in this corpus, we get fewer nouns than them.

When they use POS bigrams, they get a 95% accuracy. For us, POS bigrams and trigrams yield an accuracy of 86.21% and 87.39% respectively.

Using their topic-independent set of 18 document-level features, they achieve an accuracy of 97%. Here, once again the punctuation n-grams and POS-punctuation n-grams, both sets of features being topic independent, gave higher results: 99.41% accuracy in all cases for punctuation bigrams and trigrams, and POS-punctuation bigrams and trigrams.

Surprisingly, the cohesive markers are competitive with 83.20% and 89.02% in this scenario in contrast with the parallel corpus.

D. EXPERIMENTS WITH BOTH IBSEN CORPORA

The results of using both Ibsen subcorpora are presented in this subsection. Here, we present two scenarios: first, training in the parallel corpus and testing on the nonparallel, and, second, the opposite operation—training in the nonparallel and testing in the parallel.

Table 7 presents the results of training in the parallel corpus and then testing on the nonparallel corpus.

In this first case, Lynch and Vogel [20] reported an accuracy of 90%. They used the 18 document-level features and found some features to be the most useful: the simple to complex sentence ratio and the average sentence length.

In principle, we would expect that by training the classifier on a parallel corpus, the classifier would be able to pick up the ''style'' of the translator. If that style is consistent, it would suffice to identify the translator in other plays.

Once again, punctuation proves useful giving us higher results than Lynch and Vogel's. In particular, the trigrams of punctuation give us a 98.84% accuracy.

TABLE 8. Accuracy results of training with the nonparallel Ibsen corpus and testing on the parallel corpus using χ^2 feature selection ($k = 25$).

Features	SVC (%)	NB (%)	DT (%)	LR (%)
word unigrams	63.27	**65.31**	55.10	61.22
word bigrams	73.47	**73.47**	59.18	57.14
word trigrams	**53.06**	**53.06**	51.02	42.86
punct unigrams	63.27	61.22	61.22	**65.31**
punct bigrams	**91.84**	85.71	89.80	83.67
punct trigrams	89.80	87.76	**91.84**	83.67
POS bigrams	**59.18**	**59.18**	57.14	51.02
POS trigrams	**59.18**	55.10	55.10	46.94
POS-punct bigrams	87.76	87.76	**97.96**	83.67
POS-punct trigrams	81.63	83.67	95.92	**100.0**
cohesive	55.10	55.10	59.18	**63.27**
cohesive w/punct	57.14	57.14	**61.22**	57.14
syntactic bigrams	**61.22**	57.14	55.10	59.18
syntactic trigrams	**61.22**	59.18	57.14	55.10

TABLE 9. Top 10 word unigrams for the translators of *Don Quixote*.

	Jarvis	Ormsby	Shelton
1	answered	n't	hath
2	priest	on	quoth
3	therefore	however	yet
4	who	has	sir
5	which	curate	that
6	you	thee	unto
7	yet	worship	although
8	has	said	doth
9	thereof	that	thou
10	worship	wherein	therefore

Table 8 presents the results of the opposite operation: training in the nonparallel and testing on the parallel.

For this set-up, Lynch and Vogel report an accuracy of 83.33% using again the 18 document-level features. Our results are higher, once again in the case of POS-punctuation n-grams. In fact, we achieve perfect accuracy with POS-punctuation trigrams.

However, in neither case nor cohesive markers nor syntactic n-grams prove as discriminatory as the punctuation n-grams or the POS-punctuation n-grams.

V. DISCUSSION

In this section, we give a detailed discussion of some of our results per corpora and then present a causal explanation of using parallel and contemporaneous corpora in this task as well as an explanation for the different results for each type of feature used in the experiments.

A. DISCUSSION OF THE RESULTS ON THE DON QUIXOTE CORPUS

Recovering the most distinctive features for each translator as explained in Subsection III-F, we can see that there is a marked difference in usage of some words between translators. For example, the most distinctive word unigram for Jarvis is ''answered''; for Ormsby is ''n't'', and for Shelton is ''hath'' (see Table 9).

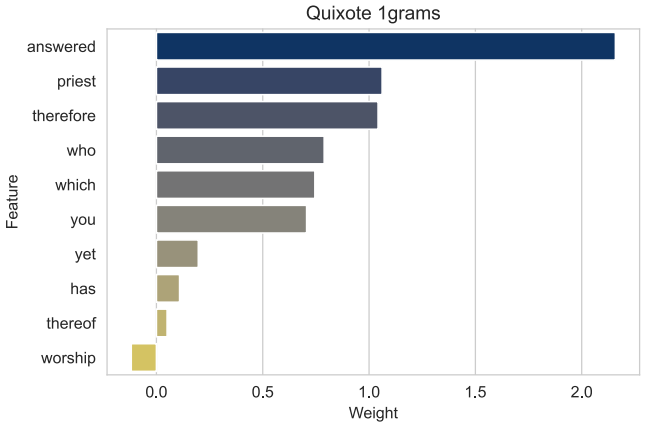

FIGURE 2. Jarvis' word unigrams "fingerprint".

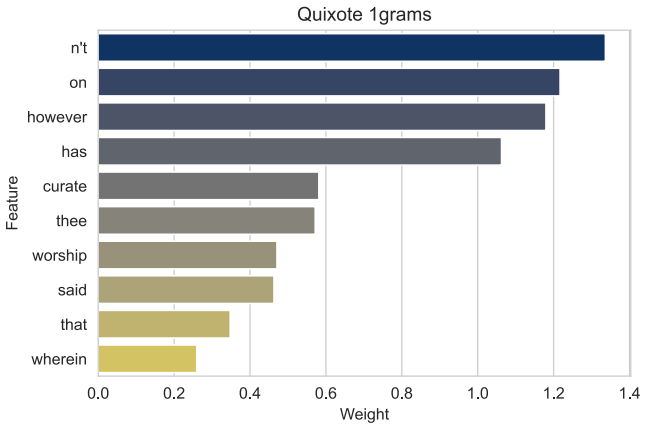

FIGURE 3. Ormsby's word unigrams "fingerprint".

The most relevant word unigram for Jarvis is, in fact, a reporting verb and if we look at the ten most relevant word unigrams for the other two translators we find the reporting verbs "said" for Ormsby and "quoth" for Shelton. This result is very important because gives support to the initial intuition of Baker of using reporting verbs as stylistic marks (see Subsection II-A).

Inspecting the rest of the ten most distinctive word unigrams, we find "therefore" for Jarvis, "however" for Ormsby, and "although" for Shelton. The three are cohesive markers, which supports our hypothesis of using them as features.

It is important to mention that some of the features are actually negative features (i.e, their presence is a negative indicator for a specific class). In other words, in some instances, a classifier did not learn ten features as *positive* indicators for a translator. For example, Fig. 2 shows a bar chart with the weights for each of the ten most relevant features for Jarvis. We can see that, in fact, "worship" is a negative indicator for Jarvis, whereas for Ormsby it is a positive indicator (see Fig 3).

Furthermore, the top ten unigrams tell us more. Some of the unigrams used by Shelton ("hath", "quoth", "doth", and "thou") are archaic forms, which is consistent with the fact that his translation is the oldest. Shelton translated the novel in the early 17th century—period of transition from Early Modern English to Modern English. The other two translations date to the 18th and 19th centuries. Thus, the classifier is effectively distinguishing Shelton's period instead of his actual personal style.

TABLE 10. Top 10 word bigrams for the translators of *Don Quixote*.

	Jarvis	Ormsby	Shelton
1	answered PROPN	at once	quoth PROPN
2	answered the	do n't	that hath
3	and therefore	i 'm	he hath
4	PROPN answered	has been	he that
5	the priest	said PROPN	unto him
6	in short	the curate	quoth the
7	your worship	your worship	by reason
8	those who	said the	i pray
9	quoth the	those who	PROPN quoth
10	PROPN quoth	that which	and therefore

TABLE 11. Top 10 syntactic bigrams for the translators of *Don Quixote*.

	Jarvis	Ormsby	Shelton
1	answered[PROPN]	at[once]	hath[he]
2	mind[a]	doubt[no]	PROPN[quoth]
3	in[short]	of[sort]	unto[him]
4	in[manner]	in[way]	hath[that]
5	priest[the]	worship[your]	pray[i]
6	answered[is]	landlord[the]	by[reason]
7	said[priest]	know[n't]	quoth[PROPN]
8	worship[your]	said[PROPN]	unto[me]
9	landlord[the]	curate[the]	said[PROPN]
10	lady[PROPN]	lady[PROPN]	said[priest]

So we can conclude that the word *n*-grams are susceptible to the period of the translation. This result again supports the initial remarks of Baker that "we could argue that the stylistic elements we identify may be explained in terms of the evolution of the target language" [5, p. 262] and is supporting evidence for using parallel and *contemporaneous* corpora for this task when using word *n*-grams or even word syntactic *n*-grams.

Regarding the word bigrams (see Table 10), again we find instances of reporting verbs in the ten most distinctive word bigrams for the three translators and some instances of cohesive markers (e.g., "in short" for Jarvis and "and therefore" for Shelton). Interestingly enough, as we already mentioned, two out of the ten most distinctive word bigrams for Shelton are "that hath" and "he hath", which are telling of the period of his translation.

When inspecting the top ten syntactic *n*-grams (see Table 11), we see an overlap with the word *n*-grams with the inclusion of reporting verbs for the three translators and archaic word forms for Shelton. The last three features for Jarvis are, in fact, negative indicators for him and positive for Ormsby, whereas "said[PROPN]" is a positive indicator for both Ormsby and Shelton, and "said[priest]" is positive for Jarvis and negative for Shelton.

TABLE 12. Top 10 punctuation trigrams for the translators of *Don Quixote*.

	Jarvis	Ormsby	Shelton
1	* * :	* * *	. ' *
2	. " -	. " "	. ' '
3	" - "	" " *	' ' *
4	- " *	" * *	, * *
5	. " *	. " *	* . '
6	* , *	* , "	* * ,
7	, * *	, " *	* , *
8	* . "	* , '	, " *
9	* : "	' ' *	* , "
10	* . '	: " *	' * *

TABLE 13. Top 10 POS-punctuation trigrams for the translators of *Don Quixote*.

	Jarvis	Ormsby	Shelton
1	" - "	. " "	NOUN . '
2	. " -	" VERB PROPN	. ' '
3	- " PRON	, " VERB	' PROPN PROPN
4	VERB : "	" " PRON	VERB , '
5	NOUN . "	PROPN , "	, ' PROPN
6	PROPN ; '	NOUN , "	' ' PRON
7	' VERB PROPN	, ' PRON	NOUN , '
8	, ' VERB	NOUN , '	PROPN , '
9	' PRON VERB	, ' VERB	" " PRON
10	' VERB DET	VERB , '	' PRON VERB

As we mentioned in Subsection IV-A, punctuation unigrams picked up some features that are not punctuation due to errors during POS-tagging (e.g., ''wind'd'', ''talkest'', ''hath''). However, for punctuation bigrams and trigrams, the features recovered do correspond to some translatorial preferences (see Table 12). We can see, for example, that Jarvis introduced dialogue using colon and double quotes (numbers 1 and 9 in Table 12), whereas Shelton made heavier use of single quotes.

The POS-punctuation *n*-grams consistently provided good results across the board. Table 13 shows the top ten POS-punctuation *n*-grams. Immediately, we see some overlap with some of the features in Table 12. Thus, POS-punctuation *n*-grams are more general than punctuation *n*-grams with the disadvantage that a POS-tagger is needed. Again, we could have some stylistic interpretation of these most distinctive features (spaCy's POS-tagger uses Universal Dependencies; see their website[8] for a description).

Finally, our proposal of using cohesive markers yields competitive results. For Jarvis, the most distinctive is ''in short''; for Ormsby is ''however'', and for Shelton is ''although''. If we extend the focus to include surrounding punctuation, the most distinctive are ''; and,'' for Jarvis; ''and'' for Ormsby, and ''' and'' for Shelton.

All these features also have a stylistic interpretation. For example, Jarvis was fond of the connector ''in short'' to summarize ideas, and Ormsby and Shelton differed in their choice of cohesive marker to denote contrast. Likewise, augmenting the markers with punctuation sheds light on the part of the sentence in which they preferred to use them. For example, Jarvis connected sentences with a conjunction but distance them with a semicolon at the same time, whereas Ormsby liked to connect them without the semicolon and Shelton used the conjunction following a direct quote.

Furthermore, in order to use them, there is no need to have a POS-tagger. A simple list of cohesive devices suffices to represent the documents, perform the translatorship attribution, and even provide a stylistic interpretation.

As we have already pointed out in Subsection II-B, El-Fiqi *et al.* [21] removed the punctuation and lemmatized the text as part of their preprocessing step. We have shown here that punctuation provides discriminating stylistic information; thus it is worth keeping it. Furthermore, we used syntactic information via syntactic *n*-grams, but showed that those features capture information relative to the period of the translation.

They mention in [21] having used the NLTK Python library for lemmatization. However, they did not specify whether they took care of the archaic verb forms. If they did not, the lemmatizer would have not identified that the base form of ''art'' is the verb ''be'', for example. Since they trained three classifiers for each pair of translators, that would explain why in [21, Tables 10 and 12] the accuracy results for Shelton *vs.* Ormsby and Shelton *vs.* Jarvis are higher than the accuracy results for Ormsby *vs.* Jarvis, which skews the average accuracy results overall.

TABLE 14. Top 10 word unigrams and bigrams for the translators of *Ghosts*.

	Archer	Sharp	Archer	Sharp
1	**'ve**	**n't**	**i 'm**	**was the**
2	**'m**	tomorrow	**do not**	**do n't**
3	**recollect**	**standing**	**i 've**	as if
4	**not**	manders	PROPN why	**it is**
5	bye	events	PROPN then	manders but
6	's	am	can not	up to
7	morrow	**getting**	not PROPN	manders and
8	softly	fear	no doubt	i am
9	'll	**because**	there 's	at all
10	dread	is	you 're	going to

B. DISCUSSION OF THE RESULTS ON THE IBSEN CORPORA

Remember that the Ibsen corpus comprises two subcorpora: one parallel and contemporaneous (i.e., the play *Ghosts*) and one nonparallel but contemporaneous (i.e., six other plays). Table 14 shows the most distinctive word unigrams and bigrams using the parallel and contemporaneous corpus with the boldface *n*-grams denoting the coincidences with the results of [20]. Our results match partially with theirs. For example, they also found the use of ''because''—a cohesive marker—by Sharp to be different from Archer and the use of contractions by Archer to be more prevalent than that of Sharp. Interesting is the finding of the use of ''tomorrow'' as one word by Sharp and as two words by Archer.

[8]http://universaldependencies.org/u/pos/

TABLE 15. Top 10 word unigrams and bigrams for the translators of the other Ibsen plays.

	Archer	Sharp	Archer	Sharp
1	looking	**community**	from me	**comes in**
2	**eyes**	be	**at him**	going to
3	with	is	PROPN with	**in from**
4	him	will	little PROPN	**the town**
5	little	**public**	PROPN softly	that the
6	then	to	**beside the**	the whole
7	softly	town	**at her**	will be
8	maid	billing	with a	it is
9	looks	her	it 's	PROPN looking
10	oh	of	the maid	the community

TABLE 16. Top 10 cohesive markers with punctuation for both Ibsen corpora.

	Archer Parallel	Archer Nonparallel	Sharp Parallel	Sharp Nonparallel
1	again!	) but	up to	. that is
2	) but	) there	, then	. but,
3	. then	) and	. of course,	; and
4	; and	) then	that is	: but
5	then	, far	, too!	: and
6	; there	then,	. indeed,	; but
7	? and	again!	, then.	: that is
8	, and–	. and	–and	. then
9	, too,	. and then	undoubtedly	) that is
10	again.	up to	again)	in all

Contrary to Lynch and Vogel, we did not find distinguishing the choosing of ''pastor'' over ''mr.'' to refer to that character. Note, however, that our POS-tagger made a mistake not identifying ''mander'' as a proper noun.

Stylistically speaking, we can say that Archer tends to use more contractions, for example. On the other hand, Sharp uses more present participles for stage directions and also in the main text—Norwegian does not have progressive tenses, so the translator has to *choose* between the progressive form in English and the simple present.

Also note that, unlike with the *Don Quixote* corpus, there are no *n*-grams product of period differences just as we expected since the translations are contemporary.

When we turn our attention to the other subcorpus, we find (see Table 15) also partially matching results with [20] (again in boldface in the table). As they point out, some topic-related *n*-grams creep into the results since the corpus is nonparallel.

For example, the *n*-grams ''community'' and ''the community'' are telling of Sharp because of the topic of a particular play. We would not say that using that specific word is part of the style of Sharp. On the other hand, it is interesting seeing ''softly'' in Tables 14 and 15 for Archer. Also, we do see that Sharp avoids using contractions (seen in the presence of the verb forms ''is'' and ''will'') meanwhile Archer favors them.

Regarding our proposed features, we found inconsistent accuracy results. Table 16 shows a little overlap between the cohesive markers with punctuation for each translator across both corpora. One possible explanation is that theater plays make less varied use of cohesive devices because it resembles spoken language more than prose does.

What we found was a consistent excellent performance on POS-punctuation *n*-grams giving us higher accuracy than the state of the art across the board. The differences between translators show stylistic preferences on the manner to give stage directions and introducing dialogues for the characters.

C. CAUSAL INTERPRETATION

We have mentioned the general and accepted recommendation of using parallel and contemporaneous corpora for identifying translators' style as suggested in [5]. The rationale

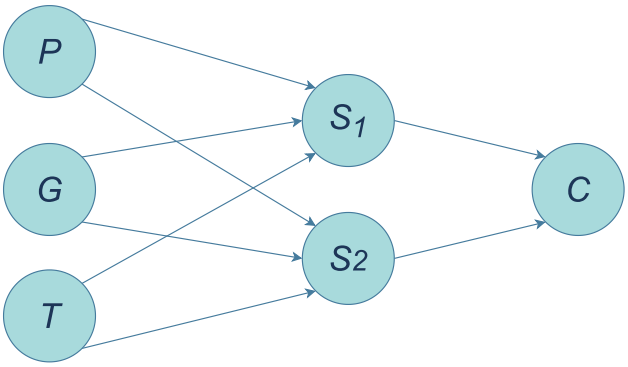

FIGURE 4. Causal directed acyclic graph (DAG) showing the "causal" influence of textual period (P), genre (G), and topic (T) on the set of features (S_i) used by translators $i = 1, 2$ employed to train a classifier (C).

behind this recommendation is that it ''enables most variables in the translation process (the author of the source text, the source and target languages, the time of publication of the original and the translation, etc.) to be held constant, so that the remaining variable, the *translator*, and his or her style, becomes the source of explanations for divergences between two translations.''(emphasis in original) [12, p. 75] as cited in [6, p. 181].

Lynch and Vogel [20] called these variables *confounding* factors and they enumerate textual period, source language, and genre as the most common. Drawing insights from causal inference (see [13], [34]), where confounding is well defined and studied, we can explain the intuition and recommendation of using parallel and contemporaneous corpora for identifying the translator's style as well as the differences in performance observed in the results of the experiments discussed in Section IV.

Fig. 4 shows a directed acyclic graph (DAG) which represents a causal graphical model. In such models, the nodes of the graph are variables and the edges represent *directed* and *causal* influences between the variables. Causality only flows along directed paths, but association can flow along any unblocked path (see [34, Ch. 2] for a full detailed description).

In Fig. 4, the period (P), genre (G), and topic (T) have a direct influence on the set (S_i) of features used for two

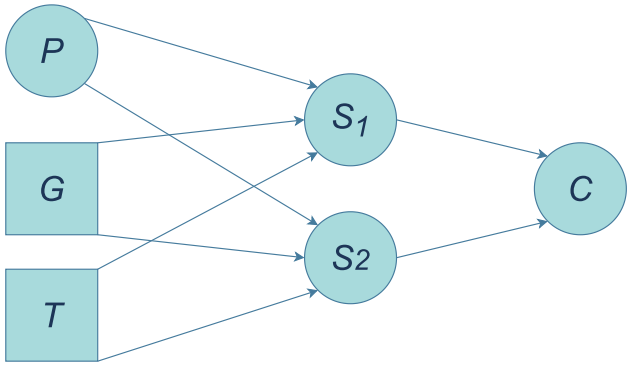

FIGURE 5. Causal DAG showing variables genre (*G*) and topic (*T*) controlled.

translators ($i = 1, 2$). (We could have added more variables such as source and target languages or original author and also more translators.) In turn, those features affect, during training, the accuracy of a classifier (C). For example, a good set of features facilitate learning a better (i.e., more accurate) classifier.

This causal approach has been used previously in the more general task of text classification (see [35]) to improve the robustness of the learned classifier. Landeiro and Culotta [35] posed the situation having a confounding variable Z directly affecting both a term vector $\mathbf{X}$ and a class label Y. Here, we differ the setting by having the confounding variables affecting the set of features (or term vector to use their phrasing) used by each translator and only through the *mediation* of those features affecting the classifier (but not directly the class label as in [35]).

When we have a parallel corpus, we are effectively controlling the effect of variables G and T (and original author and source and target languages had we added those to the DAG). We represent this control in the DAG with squares in the nodes in Fig. 5. This control within levels of G and T blocks the "backdoor path" from variables S_i to C through G and T. However, there is still one backdoor path open through P, which is enough to have a bias in the learned classifier.

We saw this effect when, in Subsection V-A, we showed that the classifier trained for Shelton with word n-grams and syntactic n-grams in the *Don Quixote* corpus picked features that depended on the period of the translation (Shelton used Early Modern English vocabulary). On the other hand, when we have parallel and *contemporaneous* corpora we are removing all these possible confounding variables (in other words, there are no more backdoor paths open between variables S_i and C).

With this explanation in mind, now it is apparent that having a parallel contemporaneous corpus is not the only way to remove bias from the classifier. For example, we can use a set of features that are not affected by the confounding variables, such is the case with POS-punctuation n-grams. Our results show that those features proved useful even when using a nonparallel corpus (see tables 6 and 8).

Just as the POS-punctuation n-grams, the cohesive markers and their punctuation are not susceptible to the period or the topic either. We saw they yielded good results in the *Don Quixote* corpus even when they are not contemporaneous. However, the results in tables 5, 6, 7, and 8 for the Ibsen corpora are inconsistent, with only good results in the nonparallel corpus.

First, we had the hypothesis that since prose and drama are different—with drama resembling more the spoken language than prose—maybe there is not as much diversity of cohesive markers usage in drama as in prose. But with the mixed results on the Ibsen corpora, the results point to either *Ghosts* being a problematic play or that the extension of the play is not enough to learn a good classifier. A more detailed stylistic or linguistic analysis of the play would be necessary to provide an explanation.

To summarize:

- The cohesive markers extended with punctuation are cheap to find, accurate, independent of period, topic, original author, source language, etc. On the downside, they appear to be sensitive to genre or size of the training data.
- Syntactic n-grams are expensive since they need a dependency parser and those are usually available only for natural languages with large resources. Additionally, they are sensitive to topic and period.
- POS-punctuation n-grams proved the most accurate across all experiments. They are not sensitive to period, genre, or topic. Furthermore, they encompass, because are more general, punctuation n-grams. On the downside, a POS-tagger is needed to build them.

VI. CONCLUSION AND FUTURE WORK

This paper addressed the understudied task of translatorship attribution, which consists in attributing a translated text to its translator. The task is related to the well-known task of authorship attribution.

The followed methodology is the one proposed by Mona Baker [5] two decades ago of using a parallel corpus of translations. In other words, using translations of the same work by different people.

Two recent works (i.e., [20] and [21]) followed this same framework and used modern machine learning techniques to approach the problem. We employed a drama corpus from [20] and one prose corpus from [21] and proposed a novel set of features on the discourse level to represent the documents in a vector space model. We also proposed to use syntactic information from dependency trees in this task.

The proposed features are cohesive devices (a list of them) by themselves and extended with their surrounding punctuation. For the syntactic information, we use syntactic n-grams [10]. We performed attribution experiments using machine learning classification algorithms with the proposed features and with word, POS, POS-punctuation, and punctuation n-grams.

The results for the cohesive markers are very positive for one corpus and mixed for the other. We hypothesize that the reason for the contrasting performance might be the difference in the genre of the texts. The translator has more ''room'' to choose a cohesive marker in prose than in drama.

However, we found the usage of POS-punctuation *n*-grams to be consistently better than the rest of the features explored in the experiments. In addition, we proposed a way of recovering the most relevant features for each translator (a sort of ''stylistic fingerprints'') from the weights of a linear classifier. In particular, we recovered the features from a logistic regression classifier trained with the entirety of the examples after having used a technique of feature selection.

Lastly, we provided a causal explanation for the rationale of using parallel and contemporaneous corpora in this task and proposed that using features that are not susceptible to confounding variables provides an alternative for not using a corpus with such characteristics. Interestingly enough, when using a parallel corpus with word and syntactic *n*-grams, the ''fingerprints'' recovered from the trained classifiers support the usage of reporting verbs and even our proposed cohesive markers as suggested in [5].

As future work, we would like to explore extending the list of cohesive devices (since our list is by no means comprehensive) and employ these features in a different corpus including several contemporaneous translations of works by different authors. We would also like to perform experiments in a non-parallel and noncontemporaneous corpus using features not susceptible to confounding variables such as original author, source and target language, topic, period of the translation, etc., and assess the performance of said features.

ACKNOWLEDGMENT

The authors would like to thank Prof. A. Gelbukh, Prof. G. Sidorov, and Prof. O. Kolesninova for their valuable inputs and comments during the development of this work. The author Christian Caballero would like to thank A. F. Silva-Martínez for her input on the causal explanation in Subsection V-C.

REFERENCES

[1] D. I. Holmes, ''Authorship attribution,'' *Comput. Hum.*, vol. 28, no. 2, pp. 87–106, 1994.
[2] P. Juola, ''Authorship attribution,'' *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, Mar. 2008.
[3] M. Koppel, J. Schler, and S. Argamon, ''Computational methods in authorship attribution,'' *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 9–26, Jan. 2009.
[4] E. Stamatatos, ''A survey of modern authorship attribution methods,'' *J. Amer. Soc Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.
[5] M. Baker, ''Towards a methodology for investigating the style of a literary translator,'' *Target. Int. J. Transl. Stud.*, vol. 12, no. 2, pp. 241–266, Dec. 2000.
[6] R. Youdale, *Using Computers in the Translation of Literary Style: Challenges and Opportunities* (Routledge Advances in Translation and Interpreting Studies), no. 42. New York, NY, USA: Routledge/Taylor & Francis, 2020.
[7] A. Akmajian, R. A. Demers, A. K. Farmer, and R. M. Harnish, *Linguistics: An Introduction to Language and Communication*, 6th ed. Cambridge, MA, USA: MIT Press, 2010.
[8] G. Ranger, *Discourse Markers: An Enunciative Approach*. Cham, Switzerland: Springer, 2018.
[9] O. Dontcheva-Navratilova, R. Jančařiková, G. Mišší́ková, and R. Povolná, *Coherence Cohesion English Discourse*. Brno, Czechia: Masarykova univerzita, 2017.
[10] G. Sidorov, *Syntactic N-Grams Computing Linguistics*. Cham, Switzerland: Springer, 2019.
[11] C. Albon, *Python Machine Learning Cookbook*. Sebastopol, CA, USA: O'Reilly Media, 2018.
[12] M. Winters, ''Modal particles explained: How modal particles creep into translations and reveal translators' styles,'' *Target. Int. J. Transl. Stud.*, vol. 21, no. 1, pp. 74–97, Aug. 2009.
[13] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton, FL, USA: CRC Press, 2020. [Online]. Available: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
[14] M. Mikhailov and M. Villikka, ''Is there such a thing as a translator's style?'' in *Proc. Corpus Linguistics*. Lancaster, U.K., 2001, pp. 378–385.
[15] R. S. Forsyth and P. W. Y. Lam, ''Found in translation: To what extent is authorial discriminability preserved by translators?'' *Literary Linguistic Comput.*, vol. 29, no. 2, pp. 199–217, Jun. 2014.
[16] S. Hedegaard and J. G. Simonsen, ''Lost in translation: Authorship attribution using frame semantics,'' in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 65–70. [Online]. Available: https://www.aclweb.org/anthology/P11-2012
[17] J. Rybicki, ''The great mystery of the (almost) invisible translator,'' *Quant. Methods Corpus-Based Transl. Stud., Practical Guide Descriptive Transl. Res.*, vol. 231, pp. 231–248, Mar. 2012.
[18] C. Lee, ''Do language combinations affect translators' stylistic visibility in translated texts?'' *Digit. Scholarship Hum.*, vol. 33, no. 3, pp. 592–603, Nov. 2017, doi: 10.1093/llc/fqx056.
[19] M. A. Covington, I. Potter, and T. Snodgrass, ''Stylometric classification of different translations of the same text into the same language,'' *Digit. Scholarship Hum.*, vol. 30, no. 3, pp. 322–325, Mar. 2014, doi: 10.1093/llc/fqu008.
[20] G. Lynch and C. Vogel, ''The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations,'' *Comput. Speech Lang.*, vol. 52, pp. 79–104, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230818301037
[21] H. El-Fiqi, E. Petraki, and H. A. Abbass, ''Network motifs for translator stylometry identification,'' *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0211809.
[22] M. Eder, J. Rybicki, and M. Kestemont, ''Stylometry with R: A package for computational text analysis,'' *R J.*, vol. 8, no. 1, p. 107, 2016.
[23] M. Honnibal *et al.*, ''Spacy: Industrial-strength natural language processing in Python,'' 2021, doi: 10.5281/zenodo.1212303.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[25] W. McKinney, ''Data structures for statistical computing in Python,'' in *Proc. 9th Python Sci. Conf.*, S. Walt and J. Millman, Eds., 2010, pp. 56–61.
[26] (Feb. 2020). *Pandas-Dev/Pandas: Pandas*. [Online]. Available: https://doi.org/10.5281/zenodo.3509134
[27] C. Molnar. (2019). *Interpretable Machine Learning*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[28] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[29] A. Géron, *Hands-on Machine Learning with Scikit-learn, Keras, and TensorFlow*. Newton, MA, USA: O'Reilly Media, 2019.
[30] H. Li, F. He, Y. Liang, and Q. Quan, ''A dividing-based many-objective evolutionary algorithm for large-scale feature selection,'' *Soft Comput.*, vol. 4, pp. 1–20, Oct. 2019.
[31] H. Li, F. He, Y. Chen, and Y. Pan, ''MLFS-CCDE: Multi-objective large-scale feature selection by cooperative coevolutionary differential evolution,'' *Memetic Comput.*, vol. 13, no. 1, pp. 1–18, Mar. 2021.
[32] M. J. Paul, ''Feature selection as causal inference: Experiments with text classification,'' in *Proc. 21st Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2017, pp. 163–172.
[33] C. D. Manning, P. Raghavan, and H. Schätze, *An Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[34] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.
[35] V. Landeiro and A. Culotta, ''Robust text classification in the presence of confounding bias,'' in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 186–193.

CHRISTIAN CABALLERO was born in Mexico City, Mexico, in 1986. He received the B.S. degree in mechatronics engineering and the M.Sc. degree in computer science from the Instituto Politécnico Nacional (IPN), Mexico City, in 2009 and 2020, respectively, where he is currently pursuing the Ph.D. degree in computer science. He also completed formal education in translation of technical and scientific texts from the Escuela Nacional de Lenguas, Lingüística y Traducción (ENALLT), Universidad Nacional Autónoma de México (UNAM), Mexico City, in 2015.

From 2012 to 2016, he was an Astrodynamicist at MEXSAT, Telecomunicaciones de México, where he was the Lead Spacecraft Engineer, from 2016 to 2019. Concurrently, he worked as a freelance translator, from 2014 to 2019. His current research interests include computational linguistics, natural language processing, machine learning, and causal inference.

HIRAM CALVO received the M.Sc. degree in computing science and engineering from the National Autonomous University of Mexico (UNAM), Mexico City, Mexico, and the Ph.D. degree in computer science from the Center for Computing Research (CIC), Instituto Politécnico Nacional (IPN), Mexico City.

He did a postdoctoral stay at the Nara Institute of Science and Technology, Japan, from 2008 to 2010. He is currently a full-time Research Professor with CIC-IPN, where he is also the Head of the Laboratory of Computational Cognitive Sciences. His doctoral thesis was on the Spanish syntax analyzer DILUCT. He is the author of more than 100 publications in the area. His research interests include computational linguistics, lexical semantics, machine learning, and psychology.

Dr. Calvo is a member of the National System of Researchers, level II, and is the President of the Mexican Association for Natural Language Processing (AMPLN). He was awarded with the Lázaro Cárdenas Prize handed by the President of Mexico, in 2006.

ILDAR BATYRSHIN (Senior Member, IEEE) received the degree from the Faculty of Control and Applied Mathematics, Moscow Physical-Technical Institute, the Ph.D. degree from the Moscow Power Engineering Institute, and the Dr.Sc. degree (Habilitation) from the Higher Attestation Committee of the Russian Federation.

He is currently a Full Professor of Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico. He is the coauthor or co-editor of more than 20 books and special volumes of journals. He presented more than 15 keynotes, Plenary talks and Tutorials on international conferences on intelligent and fuzzy systems. He is a fellow of the IFSA, a Level 3 (Highest) Researcher of the National System of Researchers (SNI) of Mexico, an Honorary Professor of Obuda University, Hungary, and an Honorary Researcher of the Republic of Tatarstan, Russia. He is the Vice-President of the Mexican Society for Artificial Intelligence and the former President of the Russian Association for Fuzzy Systems and Soft Computing.

Dr. Batyrshin is a member of the NAFIPS Board of Directors. He served as the program or organizing committee chair for more than ten international conferences on AI, CI, SC, and DM.