Post-Translation Authorship Attribution Using Stylometric Analysis

Anonymous ACL submission

Abstract

In this project, we use stylometric analysis to evaluate the extent to which a translator is able to convey the original author's style. We utilize a two-model approach, with translator and author supervised classification models, to which we pass in a continuous stylometric feature vector for each excerpt selected from our dataset. Classifiers are trained and tuned using a development set, before accuracy is generated for three chosen target texts. Our results lead us to the conclusion that of the chosen texts, Park Jung-So's Korean translation of *The Adventures of Huckleberry Finn* best preserves the stylistic qualities of its original English text.

1 Introduction

Authorship attribution and analysis is a hot topic in modern natural language processing, but is largely centred around the English language, and works written exclusively in English. While a variety of recent methods have successfully distinguished writings from different authors using lexical, syntactic and semantic information (Wu et al., 2021), this issue becomes much more complex and challenging when taken to the world of translated works (Caballero et al., 2021). Due to major differences in syntax between various languages, the task of translation involves much more than finding like-for-like word replacements, resulting in fundamentally different text structures.

Novel translation is a critical task in the modern world, allowing written works to reach a much greater audience outside of their original language. Though the primary goal of this task is to convey the semantics of the initial creation, translators also aim to portray its writing style outside of the realm of meaning, preserving qualities such as readability, lexical richness, syntactic cues and sentence structure (Piccioni, 2018). This paper aims to evaluate a small group of translated works from distinct translators on their ability to communicate these stylometric properties from their first version. Furthermore, experiments are performed on different language texts to analyze how a language's lexical vocabulary, grammar and other written rules might influence its ability to express these characteristics.

To accomplish this task, we execute a three-step process. First, recognized quantitative stylometric features are extracted from several excerpts of each of our target writings as well as training and development datasets. Next, classification models are trained and optimized to identify both the original author of each target text, as well as its translator using these features. Finally, our selected works are passed into these models, and the output predictions are interpreted and analyzed. Based on the results, we will comment on how faithful each target work is to its original author's qualities, and how much of an independent style the translator might have. Rather than making a concrete judgement on whether or not a text conveys writing style, which is a subjective metric, we compare and contrast the performance of both the translations and our classifiers.

2 Related Works

Wu et al. (2021) perform a similar task in their research, as they attempt to identify authorship through the analysis of stylistic, syntactic and semantic features. However, the authors use a multi-channel self-attention network (MSCAN) architecture (Wu et al., 2021) that takes as input text sequences and is able to extract these properties on its own. Our method involves explicitly calculating numerical features for each excerpt before passing a continuous labelled vector to a linear classifier. In addition, while we aim to attribute translated texts, Wu et al. perform experiments on original works.

Our preprocessing technique largely builds on the ideas from Elahi and Muneer (2018)'s study on identifying writing styles, where the researchers compute stylometric vectors for fragments of text

using recognized algorithms for computing readability and richness scores. They then use principal component analysis paired with an unsupervised method (clustering) to classify data (Elahi and Muneer, 2018). Conversely, our work follows a supervised approach with labelled authors and translators, also incorporating syntactic features based on part-of-speech tagging on top of this. Once again, their end goal differs from ours as they aim to classify styles themselves instead of recognizing authorship like we do.

Caballero et al. (2021) have the same motivation as we do, performing analysis on translated texts and linking them to their translators. The technique used involves passing vectors of n-gram embeddings of words, punctuation as well as POS tags into linear classifiers (Caballero et al., 2021). They recognize the challenge that this task poses when such writings carry cues reflecting the language's grammar and the original author's style. However, in our study, the latter is something we aim to identify, along with ascribing the translator. Furthermore, while the machine learning models we used were similar, the preprocessing steps and input data were vastly different in our project, as was the case for Wu et al. (2021).

3 Method

3.1 Data collection

Our research objective requires training both translator and author identification models. For this we need that for each language, the testing samples for translator models and author models are the same, in order to draw conclusions about one particular work. We also required, for each translator, a variety of works by different authors, with the aim of identifying a translator's style as opposed to a particular author's style. To our knowledge, no dataset with these characteristics has been produced before. We thus decided to compile the *Translators and Authors* dataset, which we publish along with this paper with the hope that it will serve future research.

In order to fulfill the above requirements we resorted to UNESCO's Index Translationum. The database provides a comprehensive view of global literary exchanges by cataloging translated books across various languages. Researchers can efficiently search for specific works or authors, filtering results by language and year to tailor their inquiries. Of particular importance is the ability to explore translator information within the entries. This feature allowed us to query by translator and narrow down our search to the more prolific translators. We assumed that the author and the translators of the works are the ones listed in the database records. Even though some studies apply authorship attribution to determining who the "real author" of a literary work is, this falls outside the scope of our research.

We chose our three languages (Spanish, Russian, Korean) based on their increasing level of language difficulty as compared to English, according to the United States Foreign Service Institute (2023). The motivation for this was the hypothesis that languages with similar syntax as English (i.e. Spanish) would be able to better conserve stylometric features, such as sentence length, punctuation density and readability, to give some examples.

Having chosen three prolific translators in each of our three languages, we picked the three test works, making sure that their authors were prolific enough that they had produced numerous other works to use as training data for the author identification models. We chose *Orlando: A biography* by Virginia Woolf and translated to Spanish by Jorge Luis Borges, *The Sirens of Titan* by Kurt Vonnegut and translated to Russian by Rita Rajt-Kovaleva and *The Adventures of Huckleberry Finn* by Mark Twain and translated to Korean by Jung-So Park .

We then proceeded to collect other translated works by each of the authors.

In total, we chose 5 works per author and 5 per translator. We trained each translator and author model with 3 positive and 3 negative works (i.e. translated by different translators) and for each one of them had an additional positive and negative work as a development set. The test set in each case was as described in the previous paragraph.

3.2 Feature Extraction and Preprocessing

We developed a class to generate a stylometric feature vector for an assigned text. When provided with an excerpt, it generates a list of 20 numerical measures of various qualities of the writing. These can be loosely categorized into four categories. The first, basic sentence structure, contains simple values such as the average sentence length in words and characters, as well as the ratio of characters that are punctuation. Next, eight lexical richness features are computed. These include Honore's statistic, the type-token ratio, Yule's characteristic

K, lexical entropy and others. Following this, using syllable parsing, readability features are extracted, including Flesch Reading Ease and the Gunning Fog index. Lastly, part-of-speech tags of prominent categories (nouns, verbs, adjectives, adverbs) are counted in the excerpt and ratios are computed. Some of these feature ideas were taken from Elahi and Muneer (2018), and the rest were derived either from formulas outlined in Ali Khodabakhsh (2015) or characteristics described in Zurini (2015).

It is important to note that we chose not to encode semantic qualities into our feature vector since we assume that the translators will accurately convey meaning. We instead focus on stylistic measures as defined in the introduction.

As a final step, we produce the matrices to be passed into the model. For each positive and negative source in the training, development and test sets, we extract 100 excerpts by choosing a random point in the text and reading 1000 characters. Each set of positive and negative vectors is labelled, then compiled into a feature matrix ready to be input into the models described in the next subsection. This ensures a consistent length of data, as well as a good variety of excerpts. Overlap between words is not an issue, since we are computing numeric features which will differ as long as all characters are not identical.

3.3 Translator and Author Identification Model Training

Having obtained our feature vectors, divided into training, development and testing for each author, it was time to train our Author and Translator identification models. A linear model was used in both cases to classify the data. Linear classifiers are computationally efficient and well-suited for high-dimensional data, in agreement with the 20-dimensional vectors we are dealing with. Their simplicity also makes them less prone to overfitting, particularly when the dataset is relatively small or noisy; in this case, this would be the equivalent of a short translated work, where some of the original style is lost in translation. Despite their simplicity, linear classifiers can achieve satisfactory results in a variety of practical scenarios, making them a pragmatic choice for our purposes.

The candidates we considered are Logistic Regression and Support Vector Classifier from the `sci-kit learn` library. After running some tests, we observed that SVC obtained a higher accuracy on the development set, which motivated us to carry our final tests using this model.

3.4 Cross Validation and Hyperparameter Tuning

We used `GridSearchCV` from the same library to perform hyper-parameter tuning. We used a `PredefinedSplit` object to make sure the validation set used by the Grid Search corresponds to our development set, obtained in the pre-processing stage. It is worth noting that the development set was distinctly created with excerpts from separate works. Therefore, the accuracy on the development set provides a reliable measure of the model's ability to generalize.

3.5 Testing

Once each of the six target models was trained and tuned, it was simply a matter of taking the feature matrix for each target novel and predicting its provenance using the corresponding author and translator model. Accuracy was then calculated based on the ground truth labels for each excerpt. Results are reported in section 4 and analyzed in the following sections.

4 Results

Author Model	**Target Work**		**Accuracy (%)**
	Novel	**Language**	
Woolf	*Orlando: A Biography*	Spanish	82.0
Vonnegut	*The Sirens of Titan*	Russian	64.0
Twain	*The Adventures of Huckleberry Finn*	Korean	92.0

Table 1: Accuracy of Author Models on Target Novels

Translator Model	**Target Work**		**Accuracy (%)**
	Novel	**Language**	
Borges	*Orlando: A Biography*	Spanish	68.0
Rajt-Kovaleva	*The Sirens of Titan*	Russian	43.0
Park	*The Adventures of Huckleberry Finn*	Korean	29.0

Table 2: Accuracy of Translator Models on Target Novels

Table 1 demonstrates the performance of the author models. Accuracies can be interpreted as the percentage of test excerpts that were correctly classified as either originating from the writer of the initial work (1) or not (0). For instance, 82% of the excerpts from Virginia Woolf's novel *Orlando: A Biography* translated to Spanish by Jorge Luis

Borges were correctly attributed to being **written by Woolf** by her associated model (along with negative test excerpts).

Table 2 demonstrates the performance of the translator models. Accuracies can be interpreted as the percentage of test excerpts that were correctly classified as either originating from the translator who wrote the translated work (1) or not (0). For instance, 68% of the excerpts from Virginia Woolf's *Orlando: A Biography* translated to Spanish by Jorge Luis Borges were correctly identified as being **translated by Borges** by his associated model (along with negative test excerpts).

There is a good amount of variance between all 6 models' reported accuracy, but one noticeable trend is that all three target works were substantially better attributed to their original author than they were to their translator.

5 Discussion & Conclusion

5.1 Analysis of Model Results

As pointed out previously, the author models outperform the translator models for each of their corresponding works. This reflects the idea mentioned in the introduction, whereby translators have a secondary goal to convey not only the semantics of a written text, but also its stylistic properties. Conversely, authors aim to create their own style within their work, resulting in more stylometric patterns occurring across writings from a common author as opposed to a common translator. This phenomenon creates a more accurate classifier.

Jorge Luis Borges seemed to produce a fairly accurate reproduction of Virginia Woolf's writing style in *Orlando: A Biography*, with 82% of test excerpts attributed to Woolf. He also seemed to have an identifiable style himself, as 68% of samples were correctly associated with him as a translator.

Rita Rajt-Kovaleva relatively preserved the stylometric features of Kurt Vonnegut in her translation of *The Sirens of Titan*, but if we consider the baseline accuracy to be 50% for a random classifier, this means that she barely outperformed it. There are several possible explanations for this. On one hand, it is possible her translation did not convey style very well and was more concerned with semantic preservation. Another possibility is that Russian as a language does not have a grammar well suited for representing English stylistic cues. Finally, and most probably, it could be that Vonnegut's work is not very conducive to being preserved in other languages. This explanation seems the most feasible since the author's style is uniquely colloquial in its use of American-centric expressions (Leighton, 1980). This reasoning is also reinforced by the fact that Rajt-Kovaleva does not seem to have an apparent style as a writer herself, based on our translator model results.

Finally, it can be said that Park Jung-So produced the most transparent translation in terms of stylometric features, with his translation of *Orlando: A Biography*. The reason for this is that his work was not only the most closely identified to its original author in our tests, but was very far from being attributed to the translator. As such, we can say that he succeeded the most in our original definition of the goal of translation, whereby semantic and stylometric features must be preserved.

Other than translator ability, one factor influencing our models' performance was the translation language's suitability for representing English stylistic cues. Based on our findings, it appears that Russian as a language is noticeably worse at preserving semantics and writing technique from English simultaneously. This could be a consequence of its vocabulary or its syntax limiting the variety in which meaning can be expressed. However, this is an ambitious conclusion to make, especially having only thoroughly analyzed one translation in the language. A potential extension to these experiments would be to use a richer dataset involving more translators for each language, providing a more concrete judgement on each language's ability to preserve stylometric measures through translation from English.

5.2 Assumptions

By training author models on translations done in the target language, we are assuming that the translations we select for training the author models are accurately conveying stylometric properties of the original works since we are passing their feature vectors (in the positive case) as real examples of measures for the given author.

Furthermore, we rely on our translations (in the negative case) not carrying a stylometric trend between themselves which the models might use to ascribe them as false. This would weaken the credibility of our models pertaining to the task of finding stylistic trends between positive cases, as the classifiers would likely be making decisions base on differences between an analogous negative dataset

instead.

5.3 Strengths

Careful steps were taken to avoid information leakage. No overlap exists between the original authors in training, development and testing sets for our translator models, and vice-versa for translators in the author models. This ensured that our classifiers were not gathering cues outside of the scope of what they were meant to be analyzing, i.e., using an original author's style to identify a translator.

Consistency was strengthened by training purely on fictional works, and using only translations from English into the target languages. While we considered using translations from different source languages, this would likely have impacted author model performance since different grammars would have a sizeable impact on pretty much every feature (syntactic ratios, sentence structure, lexical richness, etc.). This variance would reduce the similarity between positive samples and likely make these classifiers difficult to train. In addition, it would have been more difficult to make statements on our target languages' capacity to convey style, as we did, if we were not comparing them on a common basis of source language.

5.4 Limitations

Due to the time allotted and resources available, the data harvested was not as expansive as we would have hoped. Collection was a time-consuming process, as we needed to verify if author-translator pairs had sufficient other works that were accessible to us in the desired language. Though hundreds of excerpts were used for training, provenance was limited to three distinct works, along with one more for development. This meant that our models may have been more reflective of the precise works, notably through novel-specific vocabulary and structure, than the writers who produced them. Ultimately, we hoped to evaluate more target works in other languages and train on similar numbers of excerpts, but selected from a higher volume of original works.

Another valid criticism is our choice to use binary classification models instead of opting for a single multi-class classifier for all authors and all translators. Because of the format in which we generated our data vectors, we decided binary was more suitable, since embedding would have been required in order to accurately encode each author in a representative feature space. However, binary decision tasks are of course going to present higher performance since in theory a random classifier achieves 50% accuracy. This means our results are inflated in comparison to a model that can identify multiple authors. A potential extension to this study, along with increasing the number of target works and languages, would be to embed vectors into a feature space that takes into account properties such as their literary genre, and maybe even the syntactic complexity of their language's grammar.

5.5 Conclusion

Using a custom dataset, we were able to attribute authorship to both original authors and translators of translated novels using stylometric feature extraction and supervised machine learning. We then analyzed the results, making conclusions about each translator's ability to carry the writing style of each work.

Based on our model results, we have determined that Park's translation best preserved stylometric qualities of its original English work, followed by the versions from Borges and finally Rajt-Kovaleva. This would also lead us to believe that Korean and Spanish are more suited to conveying stylistic properties of language than Russian; however, more data is necessary to support this conclusion.

6 Statement of Contributions

All design choices and planning of tasks was done interactively as a pair. With regards to data collection, author-translator pairs and works were decided together, while Alberto gathered the necessary raw data. On the code side, Taz performed the stylometric analysis by implementing algorithms for the various measures, while Alberto took the preprocessed data, vectorized and trained both sets of models. Tuning and validation were done in a team environment. Report writing was divided between both students, each wrote about the tasks they completed, and worked together to complete the analysis, references and formatting.

7 Link to GitHub Repository

https://github.com/PythonSemicolon/550FinalProject

References

Authorship Attribution with Python | AICBT — aicbt.com. https://aicbt.com/authorship-attribution/. [Accessed 19-12-2023].

a. jdevera/pylabeador.

a. Koichiyasuoka/bert-base-russian-upos.

Koichiyasuoka/esupar.

b. Koziev/rusyllab.

b. Plantl-gob-es/roberta-large-bne-capitel-pos.

Three percent translation database.

Unesco index translationum.

Ekrem Guner Cenk Demiroglu Ali Khodabakhsh, Fatih Yesil. 2015. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, (9). [Accessed 19-12-2023].

Douglas Bagnall. 2016. Author identification using multi-headed recurrent neural networks.

J.M. Barrie. 1911. *Peter and Wendy*.

Christian Caballero, Hiram Calvo, and Ildar Batyrshin. 2021. On explainable features for translatorship attribution: Unveiling the translator's style with causality. *IEEE Access*, 9:93195–93208.

Lewis Carroll. 1865. *Alice in Wonderland*.

Philip K. Dick. 1968. *Do Androids Dream of Electric Sheep?*

Hassaan Elahi and Haris Muneer. 2018. Identifying different writing styles in a document intrinsically using stylometric analysis. [Accessed 19-12-2023].

William Faulkner. 1932. *Light in August*.

William Faulkner. 1948. *Las palmeras salvajes*. Editorial Sur.

William Faulkner. 1957. *The Town*.

F. Scott Fitzgerald. 1925. *The Great Gatsby*.

John Galsworthy. 1924. *The White Monkey*.

Charles Dale Hollingsworth. 2012. Syntactic stylometry: Using sentence structure for authorship attribution. [Accessed 19-12-2023].

Foreign Service Institute. 2023. Foreign language training - united states department of state.

Alexander H. Key. 1970. *The Incredible Tide*.

Lauren G. Leighton. 1980. Rita rajt-kovaleva's vonnegut: A review article. *The Slavic and East European Journal*, 24(4):412–419.

Pekka Lintunen and Mari Makila. 2014. Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language*, (12). [Accessed 19-12-2023].

Jody Lisberger. 2016. Creating the illusion of action: Verb density. [Accessed 19-12-2023].

Chao Lu, Yi Bu, Jie Wang, Ying Ding, Vetle Torvik, Matthew Schnaars, and Chengzhi Zhang. 2018. Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, 70.

Herman Melville. 2018. *Bartleby, the Scrivener*. la marca editora, Buenos Aires.

Brian C. Muraresku. 2020. *The Immortality Key*.

Benjamin Murauer and Gunther Specht. 2021. Small-scale cross-language authorship attribution on social media comments. [Accessed 19-12-2023].

George Orwell. 1949. *1984*.

Stephanie Payette Piccioni. 2018. The Goal of the Translator: Questioning the Impossibility of a Perfect Translation — cedar.wwu.edu. [Accessed 19-12-2023].

Edgar Allan Poe. 1845. *The Purloined Letter*. La Maquina del Tiempo.

Rick Riordan. 2010. *The Lightning Thief*.

J.K. Rowling. 1997. *Harry Potter and the Philosopher's Stone*.

J.D. Salinger. 1951. *The Catcher in the Rye*.

Mark Twain. 1876. *Tom Sawyer*.

Mark Twain. 1882. *The Prince and the Pauper*.

Mark Twain. 1884. *Adventures of Huckleberry Finn*.

Mark Twain. 1889. *A Connecticut Yankee in King Arthur's Court*.

Mark Twain. 1893. *The Tragedy of Pudd'nhead Wilson*.

Kurt Vonnegut. 1952. *Player Piano*.

Kurt Vonnegut. 1959. *The Sirens of Titan*.

Kurt Vonnegut. 1962. *Mother Night*.

Kurt Vonnegut. 1963. *Cat's Cradle*.

Kurt Vonnegut. 1973. *Breakfast of Champions*.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. [Accessed 19-12-2023].

Walt Whitman. 1855. *Hojas de hierba*.

Virginia Woolf. 1925a. *The Common Reader*.

Virginia Woolf. 1925b. *Mrs Dalloway*.

Virginia Woolf. 1929. *A Room of One's Own*.

Virginia Woolf. 1937. *Orlando: Una biografía*.

Haiyan Wu, Zhiqiang Zhang, and Qingfeng Wu. 2021. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*, 111:107815.

Yanhui Zhang and Weiping Wu. 2021. How effective are lexical richness measures for differentiations of vocabulary proficiency? a comprehensive examination with clustering analysis. *Language Testing in Asia*, (15). [Accessed 19-12-2023].

Madalina Zurini. 2015. Stylometry metrics selection for creating a model for evaluating the writing style of authors according to their cultural orientation. *Informatica Economica*, 19.