Authorship attribution in translated texts: a stylometric approach to translator style

Ana Pagano
UFMG / Brazil
anapagano.ufmg@gmail.com

Carlos Perini
UFMG / Brazil
perini@ufmg.br

Evandro Cunha
UFMG / Brazil
cunhae@ufmg.br

Adriana Pagano
UFMG / Brazil
apagano@ufmg.br

Abstract

This paper presents an exploratory study of stylometry for authorship attribution in translated texts based on characteristics of translator style. The study aimed to assess to what extent stylometric methods were successful in attributing a translated text to a particular translator (classification task) and clustering translated texts by the same translator (clustering task). To that end, a corpus of eighteen texts was compiled, including novels and short stories, originally written in English and translated into Brazilian Portuguese. Six different translators were included, each of them having authored three translated texts. The classification task was performed using a Python script for three stylometric methods: Mendenhall's Characteristic Curves of Composition; Kilgarriff's Chi-Squared Method; and Burrows' Delta Method. The clustering task was carried out in the R programming environment using various parameters available in the `stylo` package. The results were partially successful, with authorship of some of the translated texts correctly attributed to their translators in both the classification and clustering tasks. The study also found that texts by translators contemporary to each other were clustered as more similar to one another and that some translated texts were clustered due to being translations of the same original text, regardless of being authored by different translators. Our findings are in line with previous stylometric studies of translated texts, which point to the original text's style as bearing an impact on both classification and clustering tasks of translated texts.

1 Introduction

Within digital humanities, stylometry has been pursued for a variety of tasks, among them author verification, plagiarism detection, author profiling or characterization, and detection of stylistic inconsistencies (Stamatatos, 2009). Additionally, disciplines such as forensic linguistics and translation studies have also resorted to stylometric methods: the former in order to explore stylistic aspects of texts that can support correct authorship attribution for forensic purposes, and the latter to investigate characteristics of the so-called "translator style" (Saldanha, 2011).

With regard to stylometric approaches to translator detection, Rybicki (2012, 2013) carried out studies of translated texts, having obtained, at least for the language pairs he considered, partially successful and inconclusive results regarding the potential of stylometry in authorship attribution to a translated text. For the English-Brazilian Portuguese language pair, no studies, to the best of our knowledge, have reported results of stylometric techniques for the investigation of translator style. This paper seeks to fill this gap by reporting on a stylometric analysis of translated literary texts. To this end, it draws on a set of translations published in Brazil by translators who were very actively engaged in translation activities in two different historical periods: between 1930 and 1955 and between 1990 and 2015. Considering that choices made by translators reveal ways in which they see their role as cultural mediators, there is a characteristic translation style of the time, which manifests itself in traces of each translator style (Baker, 2000).

The aim of this study is thus to contribute to digital humanities and to the disciplinary field of translation studies by (i) exploring the concept of translator style from the perspective of stylometry and (ii) inquiring into how author attribution based on stylometry can be applied to translated texts in the English-Brazilian Portuguese language pair.

The remainder of this paper is organized as follows. Section 2 presents a review of the main concepts that guide our analysis. Section 3 presents the methodology for corpus compilation and analysis. In Section 4 we report results obtained. Section 5 discusses our results with respect to the available

literature. Finally, Section 6 draws conclusions from our study and presents the limitations of our work as well as perspectives for further research. Sources for our corpus and the bibliography supporting our study are provided in the References.

2 Review

2.1 Translator style

According to Baker (2000), a translator's style is a "*a kind of thumb-print*" that can be mapped based on non-linguistic and linguistic characteristics. Non-linguistic characteristics are the choices made by a translator regarding the type of text and the authors a translator decides or agrees to translate. Linguistic characteristics are recurring grammatical and lexical choices, which may or may not be conscious on the part of the translator.

Saldanha (2011) further complements Baker's definition by adding that style is a set of recurring patterns in different texts translated by the same translator, which occur regardless of the style of the original text. Saldanha (2014) also highlights the connection between the concept of translator style and that of "audience design" (Mason, 2000), which posits that the way in which translators see their role as cultural mediators and represent their readers has an impact on their translation choices, which contribute to characterize their style.

Both Baker (2000) and Saldanha (2011, 2014) use corpus linguistics concepts, such as word frequency, collocations and keywords in context, to study translator style.

2.2 Stylometry

Stylometry is an established field that explores the style of texts from a quantitative perspective, generally through computational methods. The assumption is that every author has a particular and consistent way of writing that can be recognized based on their use of lexical words (nouns, verbs, adjectives, adverbs), grammatical words (articles, prepositions, conjunctions), length of sentences used, use of punctuation marks, among other features (Stamatatos, 2009).

The first quantitative studies of style date back to the 19th and early 20th centuries (Mendenhall, 1887; Yule, 1939, 1944; Zipf, 1932). In subsequent decades, several studies were developed with a view to ascertaining which textual characteristics were most productive in author attribution of a work within the scope of what we today call stylometry (Holmes, 1994, 1998).

With developments in computer processing, stylometry (or computational stylometry) began to be approached by natural language processing (NLP) as a form of natural language understanding, with a view to extracting both knowledge and metaknowledge about texts, the latter referring to knowledge about the author of the text as a kind of psychological and sociological profile (Daelemans, 2013).

According to Laramée (2018), the lexicon that a person uses is a particular characteristic of each human being: some authors make use of a more limited lexicon than others. A writer, especially a literary one, is expected to have a more extensive and fine-grained vocabulary. However, a renowned writer such as Ernest Hemingway is frequently referred to as an author who makes use of a relatively small number of unique words when writing (Rice, 2016). This does not implicate lesser value in terms of his writing; it is deemed a matter of style.

In stylometric studies, unlike in NLP approaches and disciplines like corpus linguistics, function words, such as articles, prepositions and conjunctions, are particularly important. Stamatatos (2009) presents a review of stylometric methods and highlights the use of function words as being important as they are "used in a largely unconscious manner by the authors, and they are topic-independent" (p. 540). For stylometric analyses, function words within a corpus of works by the same author tend to vary less than lexical words.

2.3 Author attribution

Within the scope of computational stylometry, Rybicki (2012) suggests that author attribution implicates a machine learning approach for a classification task. In this process,

> the traceable differences between texts in a corpus are first used to produce a set of rules – a classifier – for discriminating authorial "uniqueness". The second step is to use the trained classifier to assign other texts samples to the authorial classes established by the classifier. (Rybicki, 2012)

For classification tasks in stylometric studies, three well established methods are explored by Laramée (2018), briefly described in the three following subsections.

2.3.1 Mendenhall's Characteristic Curves of Composition

Mendenhall (1887) proposed characterizing an author's style by a curve that expresses the distribution of the length of the words used. This is accounted for by the idea that in an author's writing, certain personal characteristics become recurrent throughout their career and these have to do with the frequency of use of short and long words. Thus, a person's writing can be characterized by counting the size of the words they use and how often this size varies.

Mendenhall compared several authors from the same historical period, counted the number of characters in each word they used and calculated the number of words with the same length. He started by counting the first 1000 words and then took random excerpts from their works. He observed that there was a pattern in word length that was repeated across different samples from the same author. Mendenhall asserted that curves generated from word sets extracted from various works by the same author will closely resemble the characteristic curve of this author.

In his proposal, a set of n words is taken from a text and, from there, a graph is created showing the frequency and size of the words in a curve.

2.3.2 Kilgarriff's Chi-Squared Method

Kilgarriff proposed using the chi-squared statistic to measure the "distance" between the lexicon used in two sets of texts. Unlike Mendenhall, whose method relied on word length distribution, Kilgarriff relies on word frequency distribution.

His method requires two corpora and selecting the n most common words in the larger corpus. He stated that the number of words to be considered is a matter not yet solved, the literature pointing to numbers between 100 and 1,000 of the most common words.

In Kilgarriff's method, the smaller the chi-squared value obtained, the more similar two texts will be and the more certain we can be that both texts were written by the same author. The assumption is that word usage patterns and a person's lexicon are very constant in an author's career.

2.3.3 Burrows' Delta Method

Burrows proposed a statistic delta value to express the distance between a text to which authorship must be attributed and a set of other texts whose authorship is already known within a corpus. Unlike Mendenhall and Kilgarriff, Burrows focuses on function word frequency and his delta is calculated by comparing the relative frequencies of function words.

The method receives this name because it measures the difference between a sample text of an author to be discovered and the other works compiled in a corpus by a known author, generating a delta value.

From this delta value, it is possible to rank candidate authors of the sample text in terms of probability of authorship. The author that is most similar will be the one whose delta has the lowest value.

2.4 Stylometry and translation

Rybicki (2012) reports a study in which he seeks to verify whether stylometric techniques are efficient to correctly attribute an author to a translated text, that is, whether translations done by the same translator are correctly identified as having the same author. Rybicki analyzes different corpora of translations of novels in two different language pairs (English-Polish; English-French). His results show that, regardless of the language pair, stylometric techniques group translated novels according to the author of the original texts instead of the translator. Instead of Burrows' Delta, Rybicki suggests using his Zeta and Iota methods, which are based, respectively, on words with intermediate frequency and the least frequent or most singular words used by a translator. Rybicki (2013) complements the results of his studies in Rybicki (2012). In studies of translated texts, stylometric methods can more successfully detect the author of original texts rather than the translator. As Rybicki highlights, the style of translated texts of the same original seems to bear similarities despite the fact that the translated texts were authored by different translators.

3 Methodology

3.1 Corpus compilation

The corpus used in our study is monolingual and comprises 18 texts translated into Brazilian Portuguese, authored by 6 Brazilian translators, each translator being the author of 3 texts.

The criteria for compiling the corpus were: (i) texts should be translations of novels or short stories originally published in English; (ii) texts should be first translations and/or retranslations into Brazilian Portuguese published in Brazil between 1930-1955 and 1990-2015; and (iii) texts

Translator name	Title and publication year of original text	Title and publication year of translated text	Label assigned	# Tokens
Monteiro Lobato	The adventures of Huckleberry Finn (1884)	As aventuras de Huck (1934)	Lobato1	82,355
	A farewell to arms (1929)	Adeus às armas (1942)	Lobato2	77,201
	The thin man (1934)	A ceia dos acusados (1936)	Lobato3	47,031
Érico Veríssimo	Of mice and men (1937)	Ratos e Homens (1940)	Verissimo1	28,470
	Point Counterpoint (1928)	Contraponto (1943)	Verissimo2	183,001
	They kill horses, don't they? (1935)	Mas não se mata cavalo? (1947)	Verissimo3	25,285
Mário Quintana	Lorde Jim (1900)	Lorde Jim (1939)	Quintana1	97,840
	God's men (1951)	Debaixo do céu (1955)	Quintana2	151,883
	Tales from Shakespeare (1807)	Contos de Shakespeare (1943)	Quintana3	83,690
Julieta Cupertino	Lorde Jim (1900)	Lorde Jim (2002)	Cupertino1	251,661
	The end of the tether (1902)	O fim das forças (2000)	Cupertino2	51,388
	Bliss and other short stories (1920)	Felicidade e outros contos (1991)	Cupertino3	34,443
Rubens Figueiredo	I married a dead man (1948)	Casei-me com um morto (1996)	Figueiredo1	64,405
	The circle (2013)	O círculo (2013)	Figueiredo2	147,033
	The thin man (1934)	O homem magro (2002)	Figueiredo3	60,670
Renato Pompeu	They kill horses, don't they? (1935)	A noite dos desesperados (2000)	Pompeu1	50,919
	No pockets in a shroud (1937)	Mortalha não tem bolso (2002)	Pompeu2	53,599
	The friends of the friends (1896); The country of the blind (1911)	Os amigos dos amigos (2004); Em terra de cego (2004)	Pompeu3	20,545
Total				1,511,419

Table 1: Corpus composition and token distribution

should make up three sets of translations by the same translator.

Monteiro Lobato, Érico Veríssimo and Mário Quintana fulfilled our criteria for the period from 1930 to 1955, whereas Rubens Figueiredo, Julieta Cupertino and Renato Pompeu met our criteria for the period from 1990 to 2015.

Two works translated by each translator were used to characterize their style in the training set, and a third one was used as a testing set for the classification task to verify whether the techniques used allowed correctly inferring author attribution based on the degree of similarity of each work in the testing set with the style of each translator as characterized on the basis of the training set. Table 1 shows the texts compiled in our corpus, their label and number of tokens[1]. The whole corpus totaled 1,511,419 tokens.

Texts were converted from their epub or pdf editions to UTF-8 encoded txt files. File preparation procedures were performed as follows: (i) assigning file name labels; (ii) clearing metadata (author name, title, title page, pagination) and additional metatext in the text; (iii) clearing symbols, spaces and blank lines. Due to pending copyright clearance for some of the texts, access to the corpus is available for research purposes upon request.

3.2 Stylometric analysis

The study comprised two stages. In the first one, a classification task was performed using a script the Python programming language. The analysis was based on the methodology presented by Laramée (2021) and included the three methods introduced in Section 2.3: (i) Mendenhall's Curves; (ii) Kilgarriff's Chi-Squared; and (iii) Burrows' Delta.

In the second stage, a clustering task was conducted using the `stylo` package in R (Eder et al., 2016). This package enables the customization of text grouping parameters, including language selection, unit consideration (token or character), n-gram size, and the establishment of minimum and maximum frequency of words. Additionally, choices such as the inclusion or exclusion of pronouns can be made based on the selected language. Clustering was executed for each parameter outlined in Table 6. In our study, texts were clustered using the various parameters and results compared as reported in the Results section.

4 Results

4.1 Classification task

Our first analysis explored Mendenhall's Curves of Composition. To assess which curves were closest to one another, a confusion matrix was generated as seen in Table 2, where, for each line, the lower the value (highlighted in red), the greater the similarity between two texts.

	Lobato	Veríssimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	0.564330	0.744281	0.448944	0.517613	**0.345221**	0.491614
Verissimo (test)	0.419614	0.674223	0.657576	0.718682	0.585002	**0.400529**
Quintana (test)	0.525755	**0.400119**	0.449070	0.517624	0.735718	1.012819
Cupertino (test)	0.324554	0.581069	0.404144	0.410844	**0.262137**	0.499184
Figueiredo (test)	0.916860	1.080957	0.754031	0.833025	**0.498228**	0.730975
Pompeu (test)	0.526212	0.715731	0.416241	0.471150	**0.218603**	0.627120

Table 2: Confusion matrix for Mendenhall's Curves of Composition method results

[1]Texts with lower number of tokens were used for testing and appear in the 'Label assigned' column with number '3'.

A heatmap was outputted as shown in Figure 1, where the closer the result to the blue shades of color, the greater the similarity.

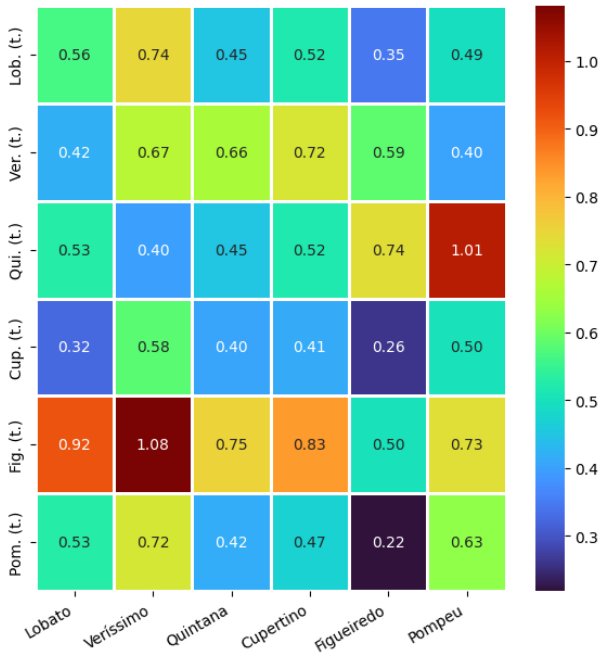

Figure 1: Heatmap of confusion matrix for Mendenhall's Curves Method results. Author names on the y axis shortened as: Lobato (test) as *Lob. (t.)*, Veríssimo (test) as *Ver. (t.)*, Quintana (test) as *Qui. (t.)*, Cupertino (test) as *Cup. (t.)*, Figueiredo (test) as *Fig. (t.)* and Pompeu (test) as *Pom. (t.)*.

As we can see, the method correctly attributed authorship to Figueiredo's text (test). The method attributed Cupertino (test) and Pompeu (test) to Figueiredo. This result could be accounted for by the fact that Cupertino, Pompeu and Figueiredo are contemporary authors (1990-2015) and hence may have more similar styles. The method also attributed Quintana (test) to Veríssimo, both authors being contemporary (1930-1955). Moreover, there are low values (higher similarity) for texts translated by translators of the same original novel, namely Érico Veríssimo and Renato Pompeu (*They shoot horses, don't they?*) and Monteiro Lobato and Rubens Figueiredo (*The thin man*).

Our second analysis implemented Kilgarriff's Chi-Squared method, with three parameters for the number of most frequent words: 500, 1000 and 5000. In this method, the lower the chi-squared result, the greater the similarity between two texts, as highlighted in Table 3.

Among all the samples, the 500-word sample was the most successful[2]. The method yielded

[2]With the parameter of 1000 and 5000 words, the method did not correctly attribute authorship to any of the translated texts.

	Lobato	Veríssimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	18601.03330	24766.78745	26889.55810	29758.62721	24399.01371	**16960.88682**
Veríssimo (test)	13144.97659	14865.49822	15389.20671	15725.75842	13834.01081	**9220.405519**
Quintana (test)	21592.12369	21491.86867	**18402.43422**	23239.12784	26385.17238	26627.86092
Cupertino (test)	14362.48120	15754.84795	16738.18427	18026.77286	14374.43532	**12695.68699**
Figueiredo (test)	19544.93223	19544.93223	29465.92648	32305.87924	22741.30033	**14912.87123**
Pompeu (test)	7218.145580	5743.387775	5766.852190	**4557.430660**	5470.142603	7324.708148

Table 3: Confusion matrix for Kilgarriff's Chi-Squared method results

greater proximity between texts by contemporary authors: Figueiredo, Pompeu and Cupertino (1990-2015), and Lobato, Quintana and Veríssimo (1930-1955), and texts translated by translators (Veríssimo and Pompeu) who translated the same original text (*They shoot horses, don't they?*).

Rank	Word	# Ocurrences
1	de	47,360
2	a	38,361
3	que	36,772
4	o	35,161
5	e	35,061
6	não	20,244
7	um	18,970
8	para	15,619
9	uma	13,924
10	se	13,679
11	com	12,732
12	ele	12,312
13	do	11,822
14	em	11,214
15	os	9,831
16	da	9,321
17	eu	8,156
18	é *(gram.)*	7,946
19	por	7,881
20	como	7,524
21	mas	7,492
22	no	7,326
23	na	6,907
24	as	6,750
25	era *(gram.)*	6,503
26	sua	5,986
27	mais	5,936
28	ela	5,806
29	você	5,566
30	seu	5,280

Table 4: Thirty most common words from all sets of texts sorted by decreasing frequency

Our third analysis explored John Burrows' Delta method. The method first extracts the thirty most frequent words in all sets, as seen in Table 4.

Table 4 shows that the most frequent words common to all sets of texts are function words, that is, pronouns, conjunctions, prepositions and articles,

including contracted forms in Portuguese (preposition plus article). Inflections of the verb "ser" *(to be)* are also part of the list.

The results of Burrows' method are displayed in Table 5. In this method, the lower the Delta score, the more similar the texts under comparison. As we can see highlighted in red, the method correctly detected the authorship of the texts translated by Quintana and Lobato. This method also yielded proximity between contemporary authors: Cupertino, Pompeu and Figueiredo, and Lobato and Quintana. Again, proximity was yielded between texts translated by Veríssimo and Pompeu, which are translations of the same original novel (*They shoot horses, don't they?*).

	Lobato	Veríssimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	**1.105159**	1.447027	1.540591	1.688335	1.564369	1.141621
Veríssimo (test)	0.944607	1.331801	1.254815	1.500599	1.209773	0.662313
Quintana (test)	1.631374	1.269202	**1.227610**	1.582967	1.675187	2.073649
Cupertino (test)	1.074294	1.172532	1.209249	1.218953	**1.067089**	1.289313
Figueiredo (test)	1.192101	1.682147	1.628778	1.704506	1.448851	0.827124
Pompeu (test)	1.438636	**1.025698**	1.174371	1.204885	1.094245	1.762203

Table 5: Confusion matrix for Burrows' Delta method results

4.2 Clustering Task

Results of the clustering task corroborate what was pointed out by Rybicki (2012, 2013). This can be seen in Figure 2, for instance, which shows a dendrogram for the results obtained for the Delta Classic distance parameter. Some of the translated texts were correctly grouped as having been authored by the same translator (Cupertino1 and Cupertino2; Lobato1 and Lobato2). Interestingly, the two translated texts authored by Cupertino were written by the same author, Joseph Conrad. In this case, clustering may be due to both translator style and original text style.

All clustering methods are shown on Table 6. For each parameter in the first column, the second column shows clustered texts which are authored by the same translator, whereas the third column shows clustered texts which are translations of the same original text. As we can see, all parameters clustered Lobatos' and Cupertino's translated texts while some of them clustered Quintana's translated texts. Lobato is the only translator whose three translated texts were clustered by some of the parameters.

In addition, all parameters clustered translated texts of the same original text: Veríssimo's and Pompeu's translations of *They shoot horses, don't they?* and Quintana's and Cupertino's translations

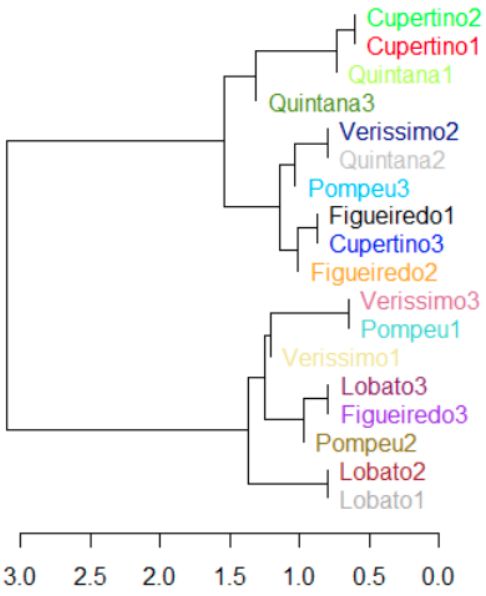

Figure 2: Classic Delta Distance Dendrogram

of *Lord Jim*, with some of them clustering Lobato's and Figueiredo's translations of *The thin man* as well.

5 Discussion

Our classification results evidence partial success, correctly classifying only one translator in the first two methods (Mendenhall and Kilgarriff), and two in Burrows'. The methods showed greater proximity between texts by translators who were contemporary, particularly within the period between 1930 and 1955, which is considered a period when translators had a more similar style (Laviosa et al., 2017).

Additionally, in all analyses, the methods associated at least one pair of texts by different translators of the same original text, which points to the impact of the original text on translation style.

Our clustering results, using different parameters, showed strong tendencies towards grouping some texts based on two aspects: (i) authorship attribution to texts by the same translator and (ii) authorship attribution to translated texts of the same original text. In other words, texts translated by the same translator cluster; so do translations of the same original by different translators. This corroborates what was pointed out by Rybicki (2012, 2013) regarding stylometric analyzes of translated texts: the algorithms are not always successful in correctly grouping texts by the same translator and there is an impact of the original text on the group-

Parameter	Clustering by translator	Clustering by original text
Classic Delta	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Cosine Delta	Lobato1 and Lobato2; Cupertino1 and Cupertino2; Quintana3 and Quintana2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Eder Delta	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Eder Simple Delta	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Entropy	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Manhattan	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Canberra	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Euclidean Distance	Lobato1, Lobato2 and Lobato3; Cupertino1 and Cupertino2.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
Cosine Distance	Lobato1, Lobato2 and Lobato3; Cupertino1 and Cupertino2; Quintana2 and Quintana3.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
MinMax	Lobato1 and Lobato2; Cupertino1 and Cupertino2; Quintana2 and Quintana3.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.

Table 6: Texts clustered according to each parameter applied

ing of translated texts.

Among the texts translated by the same translator that were most successfully grouped are the texts translated by Monteiro Lobato (the only translator whose three translated works were grouped by some of the parameters), followed by Mário Quintana and Julieta Cupertino. These results may indicate more marked style traits in these translators than in Rubens Figueiredo, Renato Pompeu and Érico Veríssimo.

The results obtained partially corroborate those obtained by Laviosa et al. (2017). The author carried out a manual analysis of characteristics noted in samples of the corpora, which grouped the translators from the periods 1930-1955 and the translators from the periods 1990-2015 into two distinct classes. The stylometric analysis carried out in our study grouped texts by the same translator, which corroborates Laviosa et al. (2017), but it also grouped translations of the same original made by translators at different times.

6 Conclusions

This study contributed to digital humanities and the disciplinary field of translation studies by pursuing research that investigated the concept of translator style from a stylometric perspective. To the best of our knowledge, the study is the first that explored this topic for the English-Brazilian Portuguese language pair.

Classification and clustering tasks were performed for authorship attribution of translated texts and the results confirmed what was observed by other stylometric studies of translated texts, with emphasis on the impact of the original text on the grouping results. We verified that stylometric methods are partially successful, both in a classification task for author attribution of a translated text and in a task of clustering texts by translator.

The main limitations of our study are: (i) variation in our corpus in terms of subgenres within the narrative genre – that is, different types of novels and short stories were used; and (ii) the predominance of male translators. These limitations

have to do with the availability and access to texts translated into Brazilian Portuguese fulfilling the inclusion criteria. Perspectives for further research include pursuing Rybicki (2012)'s suggestion to use his Zeta and Iota methods to verify whether words with intermediate frequency and less frequent or more singular words used by a translator could be indicators with more potential for author attribution of a translation.

7 Acknowledgments

The authors would like to thank three anonymous reviewers for their valuable comments. Adriana S. Pagano holds a research productivity grant awarded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6).

References

Mona Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target*, 12(2):241–266.

Pearl Buck. 1955. *Debaixo do Céu*. Livros do Brasil, Lisboa. Transl. by Mário Quintana. *Corpus Reference*.

Joseph Conrad. 1971. *Lorde Jim*. Globo, Porto Alegre. Transl. by Mário Quintana. *Corpus Reference*.

Joseph Conrad. 2000. *O Fim das Forças*. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.

Joseph Conrad. 2002. *Lord Jim; um romance*. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.

Walter Daelemans. 2013. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, pages 451–462, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with r: A package for computational text analysis. *R J.*, 8(1):107.

Dave Eggers. 2013. *O Círculo*. Companhia das Letras, São Paulo. Transl. by Rubens Figueiredo. *Corpus Reference*.

Dashiell Hammett. 1984. *A Ceia dos Acusados*. Abril, São Paulo. Transl. by Monteiro Lobato. *Corpus Reference*.

Dashiell Hammett. 2002. *O Homem Magro*. Companhia das Letras, São Paulo. Transl. by Rubens Fiqueiredo. *Corpus Reference*.

Ernest Hemingway. 2013. *Adeus às Armas*. Bertrand Brasil, Rio de Janeiro. Transl. by Monteiro Lobato. *Corpus Reference*.

David I. Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Aldous Huxley. 2014. *Contraponto*, 7th edition. Globo, São Paulo. Transl. by Érico Veríssimo. *Corpus Reference*.

William Irish. 1996. *Casei-me com um Morto*. Companhia das Letras, São Paulo. Transl. by Rubens Figueiredo. *Corpus Reference*.

Henry James. 2004. Os amigos dos amigos. In Italo Calvino, editor, *Contos Fantásticos do Século XIX*, pages 600–640. Companhia das Letras. Transl. by Renato Pompeu. *Corpus Reference*.

Charles Lamb and Mary Lamb. 2013. *Contos de Shakespeare*, 8th edition. Globo, São Paulo. Transl. by Mário Quintana. *Corpus Reference*.

François Dominic Laramée. 2021. Introdução à estilometria com Python. *Programming Historian em português*, (1).

François Dominic Laramée. 2018. Introduction to stylometry with Python. *Programming Historian*, (7).

Sara Laviosa, Adriana Pagano, Hannu Kemppanen, and Meng Ji. 2017. *A Contextual Approach to Translation Equivalence*, pages 73–127. Springer Singapore, Singapore.

Katherine Mansfield. 2000. *Felicidade e Outros Contos*, 3rd edition. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.

Ian Mason. 2000. Audience design in translating. *The Translator*, 6(1):1–22.

Horace McCoy. 1982. *Mas Não Se Mata Cavalo?* Abril Cultural, São Paulo. Transl. by Érico Veríssimo. *Corpus Reference*.

Horace McCoy. 2000. *A Noite dos Desesperados*. Sá Editora, São Paulo. Transl. by Renato Pompeu. *Corpus Reference*.

Horace McCoy. 2002. *Mortalha Não Tem Bolso*. Sá Editora, São Paulo. Transl. by Renato Pompeu. *Corpus Reference*.

Thomas C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214S):237–249.

Justin Rice. 2016. What makes Hemingway Hemingway? A statistical analysis of the data behind Hemingway's style. *LitCharts*.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator: Stylometry in translation. In *Quantitative Methods in Corpus-Based Translation Studies*, page 231–248. John Benjamins Publishing Company, Amsterdam.

Jan Rybicki. 2013. The translator's other invisibility: stylometry in translation. SLE 2013 Annual Meeting. Croatia, Split University.

Gabriela Saldanha. 2011. Translator style. *The Translator*, 17(1):25–50.

Gabriela Saldanha. 2014. *Style in, and of, Translation*, chapter 7. John Wiley & Sons.

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

John Steinbeck. *Ratos e Homens*. Livros do Brasil, Lisboa. Transl. by Érico Veríssimo. *Corpus Reference*.

Mark Twain. 2005. *As Aventuras de Huck*. Companhia Editora Nacional, São Paulo. Transl. by Monteiro Lobato. *Corpus Reference*.

H. G. Wells. 2004. Em terra de cego. In Italo Calvino, editor, *Contos Fantásticos do Século XIX*, pages 687–722. Companhia das Letras. Transl. by Renato Pompeu. *Corpus Reference*.

George Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390.

George Udny Yule. 1944. *The Statistical Study of Literary Vocabulary*. CUP Archive.

George Kingsley Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge/London.