Identifying Characters' Lines in Original and Translated Plays
The case of Golden and Horan's *Class*

Ettore Galletti
D. Interpreting and Translation
Università di Bologna, Forlì, Italy
ettore.galletti@studio.unibo.it

Alberto Barrón-Cedeño
D. Interpreting and Translation
Università di Bologna, Forlì, Italy
a.barron@unibo.it

Abstract

The aim of this project is to solve a multiclass classification problem in order to correctly identify the five characters (plus the stage directions) of the play *Class* through the stylometric analysis of each character's dialogue line. The classification is performed both for the original (EN) and the translated (IT) play in order to understand whether the same differentiation between characters is maintained or not.
First, a series of stylistic features are obtained from the corpus which contains every dialogue line of the play associated to its speaker. Then, the resulting dataset is passed through different classification models. This process is repeated three times for each language by grouping the characters in different classes: single characters, male and female, adults and children. All the models perform with similar results both for the original and the translated play, thus suggesting that the differentiation between characters does not change. However, the results are not optimal, thus suggesting the difficulty of the task and a series of problems that might be addressed in future researches.

1 Introduction

The success of a theatre play is determined not only by stage directors and actors, but also by the strength of its dialogues. Each character has its own voice and sometimes a playwright succeeds in giving each of them a strong unique voice, other times their voices are similar. Then, when a play is brought to another country, the main aspect of a successful theatrical translation is thought to be capturing the same linguistic registers (Zatlin, 2005). In this project, the aim is to verify whether the linguistic registers of *Class* are different for each character (or for different group of characters based on age and gender) and whether the differentiation is maintained in the translation in order to also provide a possible objective evaluation of a translation.
In order to do so, a corpus containing each dialogue line with its correspondent speaker was created, stage directions were also included under the name *None*. A series of stylistic features were then extracted and standardized (see Corpus). PCA and t-sne were then applied to the dataset in order to visualize possible clusters and, for the same reason, couple of features were graphically compared. Single features applied to each character were then visualized to again graphically see possible clusters. Since the corpus was highly imbalanced, a data augmentation technique was applied to the training set before passing it through the classification models. Finally, five different classifiers were used to solve the multiclass classification problem. The same process was repeated three times for each language: one with six classes (five characters + stage directions), one with three classes (male, female, stage directions), one with three other classes (adults, children, stage directions)
The rest of the report is distributed as follows. Section 2 describes the background studies in fictional voices identification. Section 3 describes the starting corpora and how they were elaborated before passing them through the classifiers. Section 4 discusses the obtained results and shows some representative examples. Section 5 presents the conclusions of this study.

2 Background

Identifying authors through the analysis of linguistic characteristics has been attempted by several scholars in the field of computational lin-

guistics and natural language processing (Ruggiero et al., 2020). However, it is still uncertain what is the best approach to do so, especially for fictional voices in literary works which are not widely studied. Moreover, different researches use different approaches and focus on different style characteristics (Muzny et al.; Madden et al.). However, stop-word frequencies, part-of-speech n-grams, and sentence lengths have been shown to be good indicators of author identity for quite a long time in this field (Stamatatos, 2009).

The present study is based mainly on the work of Vishnubhotla et al. (2019) who tried to find whether the voices of fictional characters in theatre plays can be distinguishable and whether canonical successful authors better manage to give their characters a unique voice. According to their study, using style markers and passing them through a classification algorithm allow to establish to what extent each character is distinguishable from the others. The style markers used in their study are both lexical and syntactic: n-gram frequencies of words and part-of-speech, dependency triples, average sentence and word lengths, type-token ratio, and proportion of functional words.

In the present study, classification is conducted on only a single play and its translation. Moreover, the focus is not to establish if different playwrights give a greater or narrower differentiation to their characters, but to find whether the differentiation between characters can be found in the play *Class* and whether its translation maintain this differentiation to the same extent.

3 Corpus

This study is divided into two different parts, the difference between them being the fact that the first uses the original play as its corpus and the second uses its translation.

The plays have been added to tsv files where at each dialogue line corresponds its speaker. In order to avoid a possible bias in the algorithms, every occurrence of the name of a character in the spoken lines has been masked with "charchar".

The two corpora contains the dialogue lines distributed among the characters as follows:

Character	Lines(EN)	Lines(IT)
McCafferty	393	395
Brian	374	374
Donna/Dana	303	304
Kaylie	115	116
Jayden	85	86
None	382	383
Total	**1652**	**1658**

The difference in the numbers between the English and the Italian version can be explained with the need to divide or insert new small dialogue lines for translation reasons.

It is already clear the imbalanced proportion of instances for each character in the corpus, and this issue will be addressed later in this session.

For each dialogue line, several features were extracted, namely average word lenght, average sentence length (words), average sentence length (characters), proportion of stopwords, type-token ratio, PoS n-grams with n = $\{1,2,3\}$. In order to reduce the high dimensionality of PoS n-grams, only those with more than 10 occurrences were kept. Then, all the features were standardized to avoid possible bias towards one or more of them.

Both the English and the Italian datasets were then studied in three different ways to solve three different multiclass classification problems: the first with six classes (the five characters and stage directions), the second and the third with three classes (male/female/stage directions and adults/children/stage directions respectively). This was done in order to verify a case scenario where the characters were not differentiated based on each fictional persona, but based mainly on more general characteristics such as gender and age.

The datasets were then divided into training (70%) and test (30%) sets maintaining the proportions of occurrences for each category. Thus, the sets were still highly imbalanced as shown in Figure 1.

For this reason, the training set was passed through an oversampling technique (SMOTE) in order to not penalize the less represented classes and to have the same number of occurrences. Figure 2 shows the distribution in the training set after applying SMOTE to it.

The resulting training set was then passed through five different algorithms: four machine learning algorithms - Logistic Regression, Random Forest, knn, SVC - and a neural network.

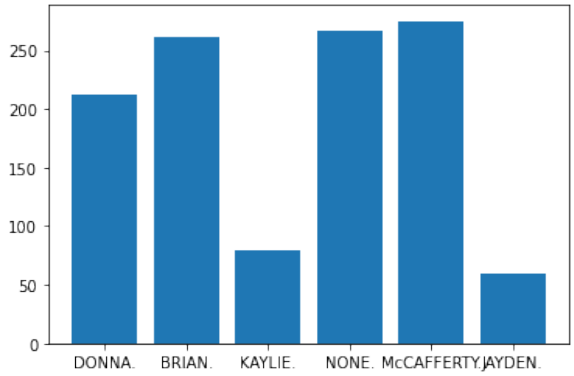

Figure 1: Distribution of lines in training set in English dataset (six classes)

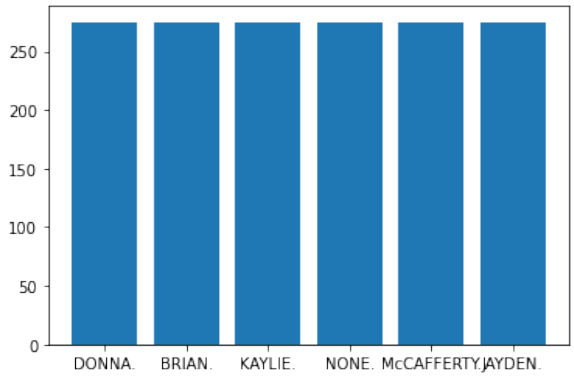

Figure 2: Distribution of lines in training set in English dataset (six classes) after SMOTE

4 Results

The results obtained are in line with what was expected, i.e. the difficulty to obtained a clear distinction between characters in almost all the case studied (six and three classes).
In the following tables the macro average of precision is reported for each classifier and for each multiclass problem (six characters, male/female/none, adults/children/none).

Algorithm	Classes	Avg (EN)	Avg (IT)
LogReg	six	0.33	0.31
	m/f/n	0.59	0.57
	a/c/n	0.61	0.60
RandFor	six	0.34	0.34
	m/f/n	0.67	0.58
	a/c/n	0.65	0.65
knn	six	0.35	0.28
	m/f/n	0.50	0.49
	a/c/n	0.54	0.61
svc	six	0.32	0.31
	m/f/n	0.61	0.56
	a/c/n	0.59	0.60
NN	six	0.41	0.53
	m/f/n	0.78	0.75
	a/c/n	0.88	0.88

For all the classifiers and for all the datasets, the class that was best identified was *None* (stage directions), probably because it does not try to reproduce a human voice as the others and it is very repetitive. For instance, if we look at the confusion matrix of the Logistic Regression for the case with six classes in Figure 3, we can clearly see this.

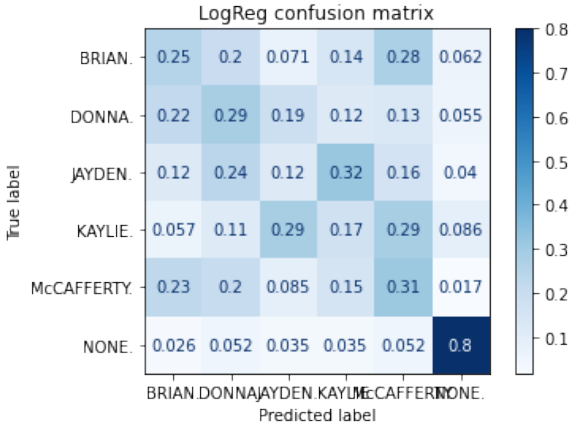

Figure 3: Logistic regression confusion matrix for EN dataset (six classes)

In general, no substantial differences in the characters' identification between English and Italian were found. The Italian datasets had, in general, a slightly lower precision.
The best performing classifier was that of the neural network for the classification of adults, children and stage directions. In general, the classification of these three classes was the best for all the algorithms both in the original and in translation. This probably mean that the authors managed to give a unique voice not to single characters but

based on the age group of the character and this differentiation was maintained in translation.
In order to visually find clusters for the characters/group of characters, PCA and t-SNE representations were made. As expected, even with these techniques no clear identifications were found as we can see in Figure 4 and Figure 5.

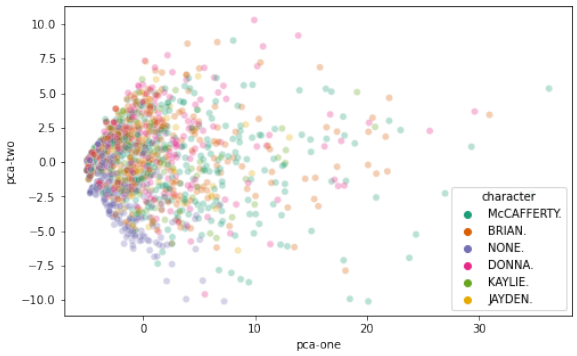

Figure 4: PCA (six classes)

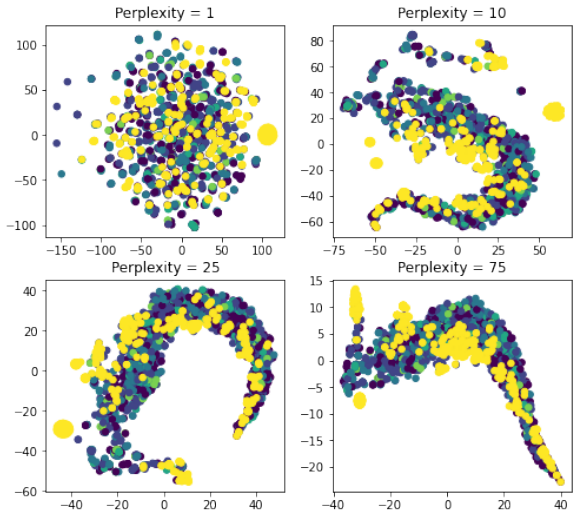

Figure 5: t-SNE with four different perplexities (six classes)

Another suggestion that the characters are similar to each other - but different from the stage directions - comes from the graphic visualization of the different features in a box plot as shown in Figure 6.

Visually comparing couple of features for each class also points out that a differentiation between characters does not emerge clearly as shown in Figure 7.

All the steps, results and the different representations for each dataset repeated for the three different classification groups can be found in the

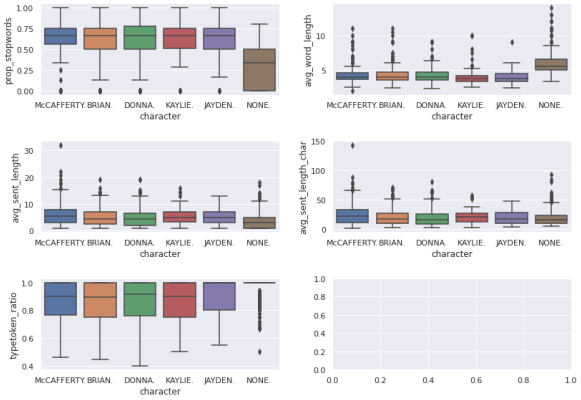

Figure 6: Box plot representation of different features (six classes)

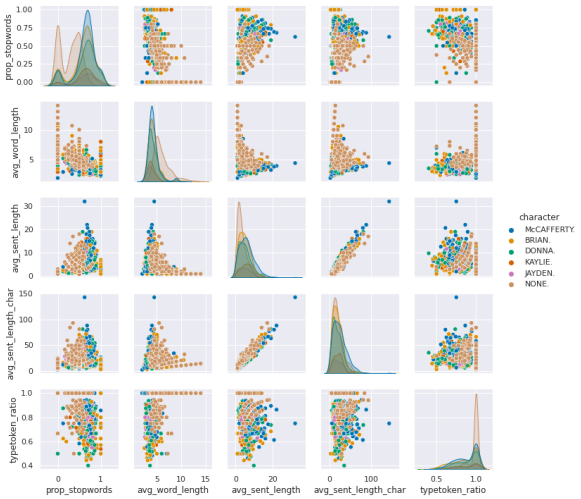

Figure 7: Couple of features compared (six classes)

Colab Notebooks here for the English corpus and here for the Italian one.

5 Conclusions

In this study, we wanted to find whether it was possible to adopt the classical features studied for authorship identification together with some of those suggested by Vishnubhotla et al. for the identification of fictional characters in a single play. The fact that only a single play was taken into account and that dialogue lines in a play are generally short probably represented an added difficulty. Nevertheless, the aim was to also find whether the identification occurs with the same precision in translation (again, in only one translation in this case) to analyse, and eventually evaluate, the translation in terms of an objective transmission of style markers.

From this study, no general assumptions can be made. Nevertheless, the task seems to be challenging, and future work is certainly needed. Future researches should probably take into account a larger corpus of plays and also evaluate more and better style markers that could have a greater impact on the subject.

References

[Madden et al.2019] Madden, Joshua and Storey, Veda C. and Baskerville, Richard 2019. *Identifying Authorship from Linguistic Text Patterns*. Digital Scholarship in the Humanities.

[Muzny et al.2017] Muzny, Grace and Algee-Hewitt, Mark and Jurafsky, Dan. 2017. *Dialogism in the novel: A computational model of the dialogic nature of narration and quotations*. Digital Scholarship in the Humanities.

[Ruggiero et al.2020] Gaetana Ruggiero and Albert Gatt and Malvina Nissim. Datasets and Models for Authorship Attribution on Italian Personal Writings 2020.

[Stamatatos2009] Efstathios Stamatatos 2009. *A survey of modern authorship attribution methods*. Journal of the American Society for Information Science and Technology, 538–556.

[Vishnubhotla et al.2019] Vishnubhotla, Krishnapriya and Hammond, Adam and Hirst, Graeme. 2019. Are Fictional Voices Distinguishable?Classifying Character Voices in Modern Drama. *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 29–34.

[Zatlin 2005] Zatlin, P. 2005. *Theatrical Translation and Film Adaptation: A Practitioner's View*. Multilingual Matters.