A Study in Post-Editing Stylometry

Ajay Krishnan **Amit Bharti** **Mehdi Amiri**

{a.krishnan.2, a.bharti.1, m.amiri.2}@student.rug.nl

University of Groningen
Groningen, The Netherlands

Abstract

Identification of an author based on his/her content is definitely gripping and fascinating. In order to achieve it, linguistic as well as behavioural key features need to be studied. In this work, we present a strategy for the identification of the translator (that we will call subjects) and modality using a random forest classifier model, while edit times were predicted using the random forest regressor model. An accuracy of 83 percent is obtained for subject identification while about 65 percent for modality identification. The right predictions of edit times were complex to be obtained for the current setup that is implemented.

The Github repository can be accessed at the following link. **https://github.com/ajaykrishnan9/stylometry_final**

1 Introduction

Stylometry is the quantitative analysis of literary style through computational distant reading strategies. It is based on the observation that authors tend to write in relatively consistent, recognizable, and unique ways (Laramée, 2022). An author may prefer to write in short sentences while others may prefer long texts consisting of numerous clauses and each author has their own distinctive vocabulary, sometimes they may be limited, and sometimes they maybe are rich. A lot of research and investigation has been made based on the linguistic and behavioral information for author identification that traces back to the late nineteenth century that was based on distributions of sentence and word lengths in literature to recent developments and practical applications in several different areas such as criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (Daniel Pavelec and Oliveira, 2010). Vocabulary richness and lexical repetition based on word frequency distributions are some features applied for several stylometric studies, along with word-class frequencies, syntactic analysis, word collocations, grammatical errors, words, sentences, clauses, and paragraph lengths.

Our main research focuses on how the three subjects differ in their technique of translation or post-editing. The translations were performed on the PET (Post-Editing Tool) platform, where a collection of the fine-grained history of key logs, edit times, and other information related to the translation process was recorded. These features are available in the train data. Our research focuses to identify the features that impact the most in identifying the subject and also how the translation modality can be predicted using the test set in which the latter and related information are masked as targets. Accurately predicting editing times from the available data, using the test set in which temporal information is masked is also part of our research.

1.1 The Team

Ajay is a CS student with a keen interest in data analysis. He worked on the technical part of the dataset, the report, as well as on the model building.

Amit is an AI student with a keen interest in solving real-world problems through technology. He worked on the temporal part of the dataset as well as on the model building and report writing.

Mehdi is an IS student. Apart from the whole research part, he specifically worked on the cognitive analysis, feature extraction, and data preprocessing provided to inject into the model.

2 Method

2.1 Terminologies

The following are some terminologies and techniques used for the prediction of the subjects, modalities, and edit times.

- **Random Forest Classifier:** This is a super-

vised machine learning algorithm utilized for classification and other tasks using decision trees. A set of decision trees are constructed by the random forest classifier from a randomly picked subset of the train data that then gathers the votes from various decision trees to determine the final prediction. (AvhijtNair, 2021)

- **Random Forest Regressor:** This is a supervised learning algorithm that uses the ensemble learning method for regression. It is an estimator that fits several decision trees on different sub-samples of the dataset. The predictive accuracy and overfitting are controlled by averaging that is performed by this regressor. (SckitLearn, 2015)

- **K-Fold Cross-Validation:** The dataset is split into 'k' number of subsets, and then k-1 subsets are employed to train the model and the final subset is kept as a validation set to test the model. The score of the model on each fold is then averaged to evaluate the performance of the model. (AskPython, 2015)

- **Confusion Matrix:** It is a NxN matrix that is used to estimate the accuracy or performance of a classification model, where N is the number of target classes. The matrix compares the true target values with those predicted values generated by the machine learning model. An overview of how well the classification model is performing and the type of errors it is making is analyzed with this confusion matrix. (Aniruddha, 2020)

- **Accuracy Score:** This function returns the score of how well the predicted labels exactly match the actual labels. The equation for accuracy is as follows:

$$\text{Accuracy} = \frac{\text{Total correct predictions}}{\text{Total predictions}} \quad (1)$$

- **Root Mean Squared Error:** This measures the average of the squares of the errors, i.e, the average squared difference between the predicted values and the true value, and considers the root of this.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x_i})^2}{N}} \quad (2)$$

For Equation 2; x_i is the true value, $\hat{x_i}$ is the predicted value, i is each row of the data and N is the number of rows.

2.2 Dataset

- **Train Data:** A small corpus of 430 sentences taken from different Wikipedia sources have been translated from English into Italian by three professional translators. In this translation, each sentence has been translated only once by each subject either by translation from scratch, post-editing of a commercial machine translation system (Google Translate), or post-editing of a multilingual research machine translation model (mBART). The train data contains the different features that have been collected from PET, on which the translations were performed.

- **Test Data:** The test data has been replicated into three test sets, where each one masking respectively the subject information, the translation modality information, and the temporal information. For example, the *subject_id* feature being masked for the subject identification test data.

2.3 Exploratory Data Analysis

Before model creation and performing the predictions to identify the subjects, modalities or edit times, the entire data is first analysed. This is done by performing the cognitive, temporal, and technical analysis based on the features provided in the train data. This is done to get an overview of the keystrokes, type of keystrokes, pace, the number and lengths of pauses taken by each subject for the translation per modality, and more. .

2.3.1 Technical Analysis

Technical analysis is performed based on the number of keystrokes used by the subjects to produce the final translation. The analysis focuses on the total number of keystrokes as well as the type of keystrokes by each subject for each translation modality. The analysis will give an insight into how much of a typing effort is needed and which type of keystrokes are hit by each subject per modality, that will differentiate the subjects. This effort also analyses the total count of words per modality for each subject.

2.3.2 Cognitive Analysis

The pauses taken by the subjects when performing the translation or post-editing are considered to measure the cognitive effort. The dependent variable can be expressed as the total count of the pauses, how long the pauses have taken, and also considering the total pause time divided by the total editing times. For the train data, the pauses of 300ms or more, as well as the pauses of 1 second or more, are considered.

2.3.3 Temporal Analysis

The temporal effort corresponds to the translation productivity that analyses the total words per hour for each subject per modality. This effort also analysis the total edit time per hour for each subject per modality.

2.4 Data Pre-Processing

After analysing the main characteristics of the dataset through visual methods, features in the dataset were prepared for the machine learning model implementation.
Firstly, categorical variables were discretized using one-hot encoding. Secondly, linguistic information was extracted from the source, machine translation, and target text through total word counts, and unique word count. Some cognitive features like the ratio from the pauses greater than 300 ms and 1000ms over edit time and word time(time spent for each word) were also formed. Part of speech (POS) tagging was then used on the above text files to get a morphological understanding from the available data. We extracted some features by counting every POS tag and differing corresponding features in both source and target texts. Lastly, edit operations such as insertion, deletion, and substitution count were added as one of the features. In addition, we considered new features as aggregated data such as the number of unique words in machine translation and target texts grouped by subject and modality. The Pre-processing is also provided for test sets considering the masked information for each.

2.5 Random Forest Modelling

Once the preprocessing is done, the available trainable data is split into two sets known as the train dataset and test dataset used for training and predicting the label from the model respectively. Since the dataset provided is limited, K-fold cross-validation is used for fine-tuning the model, hence the validation dataset is not required.

3 Experimental Setup

The different datasets are loaded for the implementation using the *Datasets* library. The *Scikit-learn* machine learning library is used for model evaluation, the train test split, random forest classifier, random forest regressor and much more. With the help of *spaCy*, a pipeline for Italian is loaded ($it\_core\_news\_sm$) for counting the POS tags on the target text.
The train data has been split into 80 percent train set and 20 percent test data set for all the strategies. The model is then trained and predicted for the labels for the split test data. As mentioned earlier, k-fold cross-validation is used for fine-tuning the model, since the dataset provided is limited, and thus the validation dataset is not required.

3.1 Subject Prediction and Effective Features Identification

For this experiment, the main goal is to use the train data to study which features are most effective in identifying a subject as the author of a translation in the test data. The subjects perform the translation of the Wikipedia sources from English into Italian. There are a total of three subjects that performed the translations. Grid search package is used for hyper tuning the random forest parameters for best results. After pre-processing and training the model, the trained model is employed to predict the subjects from the subject masked test data (*test_mask_subject*).
In order to evaluate the accuracy of the model on the subject masked test data, the true subject labels are extracted from the modality masked data set, as this dataset contains the subject information. Classification accuracy and prediction of the trained model are evaluated using confusion matrix and accuracy score.

3.2 Translation Modality Prediction

For this experiment, the main goal is to predict the translation modality, using the test set in which the latter and related information are masked as targets. Modality is the type of translation that is performed by the subjects, i.e, translation from scratch, Post-editing of a commercial MT system (Google Translate) and Post-editing of a multilingual research MT model (mBART). Also, for this prediction strategy, grid search package is used for hyper tuning the random forest parameters for best results. After pre-processing and training the

model, the trained model is employed to predict the modality from the modality masked test data (*test_mask_modality*). In this test data, along with the modality feature, features like machine translation text, the number of post-editing insertions, deletions, substitutions, and shifts are absent. Sentence level scores like BLEU, chrF, TER along with edit operation features are absent as well.
In order to evaluate the predicted modality of the model on the modality masked test data, the true modality labels are extracted from the subject masked data set, as this dataset contains the modality information. The test data is also pre-processed the same way the training data was done. Classification accuracy and prediction of the trained model are evaluated using confusion matrix and accuracy score.

3.2.1 Modality Prediction with Machine Translation as Single Category

For this experiment part of the modality prediction, both post-editing settings, i.e the mBART and Google translate modalities, are considered as a single category while the translation from scratch is another category.

3.3 Edit Time Prediction

For this experiment, the main goal is to predict the editing times from the available data, using the test set in which temporal information is masked. The editing time refers to the total editing time taken by the subjects for the translations, in seconds. Compared to the subject and modality prediction models that used the random forest classifier, the editing time prediction models use a random forest regressor. After pre-processing and training the model, the trained model is employed to predict the edit times from the edit time masked test data (*test_mask_time*). In this test data, along with the edit time feature, length and number of pauses for 300 ms and 1 second are absent as well. Prediction of the trained model is evaluated using the root mean square error.

4 Results

4.1 Data Analysis Results

As mentioned earlier, before creating the model and performing the predictions, the train data is first analysed by performing the cognitive, temporal, and technical analysis. The modality of the translation task values is referred to as the following: *ht* (translation from scratch), *pe1* (post-editing Google Translate translations), *pe2* (post-editing mBART translations), and the three subjects performing the translation from scratch or post-editing task are referred to as *t1*, *t2* or *t3*.

4.1.1 Technical Analysis

For technical analysis, a graph that illustrates the total number of keystrokes (Figure 4) as well as the type of keystrokes by each subject for each translation modality is plotted (Figure 5, Figure 6, Figure 7). The type of keystrokes referred to hear are the content keys, navigation keys, and erase keys. For this analysis, a plot that illustrates total count of words per modality for each subject is also plotted. (Figure 8)

4.1.2 Cognitive Analysis

For cognitive analysis, graphs that illustrates the ratio of the total pausing time with respect to the total editing time for each subject per modality is plotted. (Figure 9, Figure 10) Graphs for total number of pauses, as well as the length of the pauses for each subject per modality is plotted. (Figure 11, Figure 12, Figure 13, Figure 14) The length of pauses above 300ms and 1 second are considered.

4.1.3 Temporal Analysis

For temporal analysis, a graph that illustrates the total words per hour for each subject per modality is plotted. (Figure 15) A graph that illustrates the total edit time per hour for each subject per modality is also plotted. (Figure 16)

4.2 Results of Subject Prediction and Effective Features Identification

Confusion matrix and accuracy score are considered to evaluate the classification accuracy and prediction of the trained model with respect to true subjects on the subject masked test dataset.

Predicted Values

True Values	ht	pe1	pe2
ht	33	7	0
pe1	10	30	0
pe2	2	1	37

Figure 1: Confusion matrix for subject prediction

When observing the confusion matrix (Figure 1) it is found that for each subject, the predicted and

true values have a very high value ie, the diagonals of the confusion matrix. Also, when determining the accuracy score of how well the predicted subjects exactly match the actual subjects, a score of 83 percent is observed.
Features that contribute the most to the model for the subject prediction are listed in Table 1.

Features	Importance
k_nav	0.235
num_annotations	0.102
word_time	0.045
n1000_ratio	0.038
edit_time	0.029

Table 1: Feature importance for subject prediction

4.3 Results of Translation Modality Prediction

Confusion matrix and accuracy score are considered to evaluate the classification accuracy and prediction of the trained model with respect to the true modalities on the modality masked test dataset.

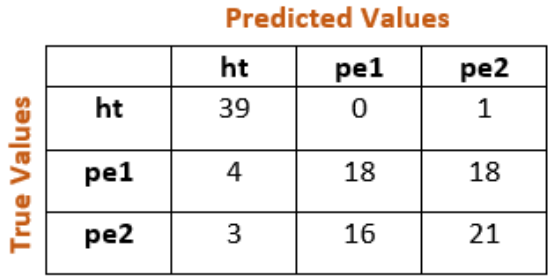

Figure 2: Confusion matrix for modality prediction

When observing the confusion matrix (Figure 2), it is found that for the two machine translation modalities, i.e, mBART, and Google Translate, the model finds it difficult the rightly distinguish between them. Also, when determining the accuracy score of how well the predicted modalities exactly match the actual modalities, a score of 65 percent is observed. Features that contribute the most to the model for the modality prediction are listed in Table 2.

Features	Importance
k_white	0.149
k_total	0.107
k_letter	0.081
k_symbol	0.071
n_pause_geq_300	0.046

Table 2: Feature importance for modality prediction

4.3.1 Results of Modality Prediction with Machine Translation as Single Category

When the post-editing settings, i.e the mBART and Google translate modalities, are considered as a single category while the translation from scratch is another category, an accuracy score of 96 percent is observed. When observing the confusion matrix (Figure 3), it is found that the model can easily classify between the machine translation and translation from scratch modalities. Features that contribute the most to the model for the this type modality prediction are listed in Table 3.

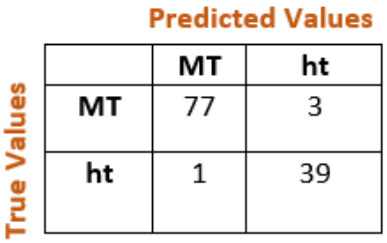

Figure 3: Confusion matrix for modality prediction with machine translation as single category

Features	Importance
k_letter	0.167
k_white	0.144
k_total	0.112
k_symbol	0.095
n300_ratio	0.073

Table 3: Feature importance for modality prediction with machine translation as single category

4.4 Results of Edit Time Prediction

As mentioned earlier, in the case of predicting edit times, the prediction of the trained model is evaluated using the root mean square error. A root mean square error of about 59.08 is observed. Features that contribute the most to the model for the edit time prediction are listed in Table 4.

Features	Importance
k_total	0.5
k_symbol	0.05
k_cut	0.04
k_copy	0.04
ratioAbnormal	0.03

Table 4: Feature importance for edit time prediction

5 Discussion

5.1 Technical Analysis

While performing technical analysis (Figure 4), it is observed that the number of keystrokes is the most, when translating the text from scratch, i.e, *ht*, as the subjects have to type more, since the translation is done from scratch and not from an already translated text, followed by *pe2* and then *pe1*. Hence, comparing between machine translation post edits, the number of keystrokes is more when performing the post-edit on mBART translations, compared to post-editing Google Translate translations.
While analyzing the type of keystrokes (content keys, erase keys, navigation keys), it is observed that the content keys are the most hit keys followed by erase keys and then navigation keys. (Figure 5, Figure 6, Figure 7) The content keys are the most hit keys by all subjects as the translations require more typing than just erasing and navigation. But as illustrated on the plotted analysis results, it is observed that the navigation keys are hit more than the erase for just the third subject i.e, *t3*, and so this is part of the analysis that clearly differentiates between users.
When analyzing the total count of words per modality for each subject, it is observed that the word counts are almost the same for every modality. And thus, this feature doesn't assist to differentiate between modalities. (Figure 8)

5.2 Cognitive Analysis

While analyzing the cognitive efforts of the subjects based on the number of pauses, the lengths of pauses, and the ratios, it is observed that in all the cases, the *ht* modality has the highest values. This is because the subjects are performing the translations from scratch for the *ht* modality and would essentially take more time for the translation than compared to post-editing the text from the machine translations. (Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, Figure 14)
When analyzing the behavior of the subjects, it is observed that subject *t3* usually takes more pauses compared to the other subjects and thus is more concerned with the editing or maybe an amateur and takes more time to think and edit. The subject *t2* usually takes less time for the translation compared to other subjects may be because the subject is a more experienced translator or would like to perform the translations in a hurry.

5.3 Temporal Analysis

While analyzing each modality for productivity, it is observed that the subjects perform the translations quickly on the *pe1* modality followed by *pe2* and then *ht*. While analyzing each subject for productivity, it is observed that subject *t2* has the highest productivity followed by *t1* and then *t3*, and thus the third subject is the slowest. This result is also consistent with previous analysis results, as it was found that subject *t3* is more of an amateur and takes more time to think and edit and thus has less productivity and that the subject *t2* is a more experienced translator. (Figure 15, Figure 16)

5.4 Subject Prediction and Effective Features Identification

When we observe the feature importance for subject prediction (Table 1), the *k_nav* which refers to the number of navigation keystrokes, contributes the most to the model for subject prediction. This result can be correlated with the technical analysis results that were observed earlier, where the navigation keys were mostly hit (or only hit) by the subject *t3* only, for all modalities, which clearly differentiated between users.
The *num_annotations* that refer to the number of times the subject focused the textbox for performing the translation of the sentence during the translation session have the second importance, followed by *word time*, *n1000_ratio* and *edit time*. These results can all be correlated with the different analysis results that were performed earlier. For example, the subject *t3* usually takes more pauses compared to the other subjects and thus is more concerned with the editing or maybe an amateur and takes more time to think and edit, compared to *t1*. The subject *t2* has the highest productivity and is thus an experienced translator.

5.5 Translation Modality Prediction

As discussed earlier, the accuracy score of how well the predicted modalities exactly match the actual modalities, a score of 65 percent is observed. Also, in the confusion matrix (Figure 2), it is found that for the two machine translation modalities, i.e, mBART, and Google Translate, the model finds it difficult the rightly distinguish between them. When we observe the feature importance for modality prediction, the *k_white* that refers to the total number of white spaces feature contributes the most to the model for modality prediction, The

feature importance is followed by the total number of keystrokes and content keystrokes. These results can be correlated with the analysis results, where it was observed that the total word count or the content keystrokes being hit is almost the same for both machine translation post edits. (Figure 8)

5.5.1 Modality Prediction with Machine Translation as Single Category

When both the post-editing setting was considered as a single category, and the translation from scratch modality as another category, it is easy for the model to classify between these two categories. This is observed in the confusion matrix. (Figure 3) As mentioned earlier, an accuracy score of 96 percent was achieved for this particular experiment. As observed in the feature importance for such modality predictions Table 3, the content keystrokes contribute a lot. This result correlates with the data analysis results that were performed earlier when there is quite a difference between the number of content keystrokes to be hit between the translations from scratch and the machine post edits.

5.6 Edit Time Prediction

The right predictions of edit time were complex to be obtained for the current setup that is implemented, which could be seen by the high value of the root-mean-square error. In the test data, along with the edit time feature, the length and number of pauses for 300 ms and 1 second are absent as well. The features that contributed the most to edit time prediction are the total keystrokes, which do not give enough information related to edit time for its correct prediction under the current implementation setup. This makes it difficult for the modal to rightly predict the edit times.

6 Conclusions

From initial exploration of the train data, it is observed that the navigation keystroke clearly differentiates between the subjects, and is also the most important feature that contributes to the model for subject prediction. It is also observed that the productivity of the subjects, as well as edit times, clearly differentiate between subjects. For modality prediction, the total content keystrokes or the count of words clearly differentiate between translation from scratch and the two machine translations. The number of pauses and length of pauses also help to distinguish between modalities. When analyzing just the subjects, we could find that the subject *t3* is the slowest and takes more time to edit and thus is an amateur and the subject *t2* is more of an experienced translator and have high productivity. When analyzing just the modalities, the translation modality *ht* takes the most time to translate, as the translation is performed from scratch, compared to the two machine translations. Whereas google translation required minimum keystrokes to edit. Therefore, google translation performs better than m-Bart while translating from English to Italian.

The following are the future researches,

- As discussed, through data analysis we saw that subject *t3* is the slowest, and different modalities have different edit times but the combination of them could give good results when the model is trained on the temporal data but mask them in the test data by using correlation matrix between temporal data and the features. This could help us to improve the edit time prediction accuracy.
- To learn in-depth about the two machine translations, i.e, mBART, and Google Translate, and how these modalities differ in their translations, will help in improving the modality predictions between these two.
- Since the train data contains only a small corpus of 430 sentences taken from different Wikipedia, the initial prediction model was implemented using SVM (Support vector machine). Later for better performance, the models were implemented using a random forest classifier and regressor. But if we have more data collected from the PET platform, then neural networks could be used for the predictions with a better performance and prediction accuracies. Although interpretability of the deep learning algorithm will be low.

References

Aniruddha. 2020. Confusion matrix for machine learning.

AskPython. 2015. K-fold cross-validation in python using sklearn.

AvhijtNair. 2021. Random forest classifier using scikit-learnn.

Edson Justino Daniel Pavelec and Luiz S. Oliveira. 2010. Author identification using stylometric features.

François Dominic Laramée. 2022. Introduction to stylometry with python.

SckitLearn. 2015. Random forest regressor.

7 Appendix

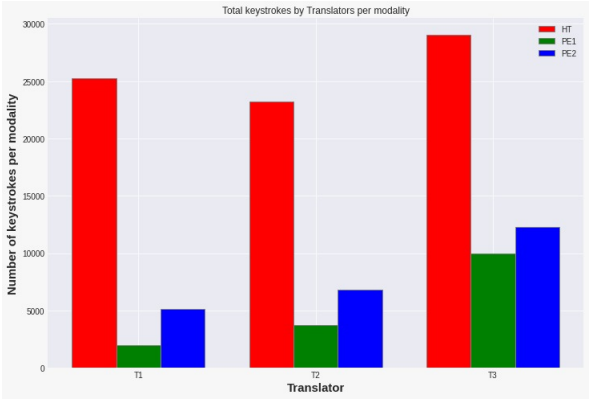

Figure 4: Total number of keystrokes for each subject per modality

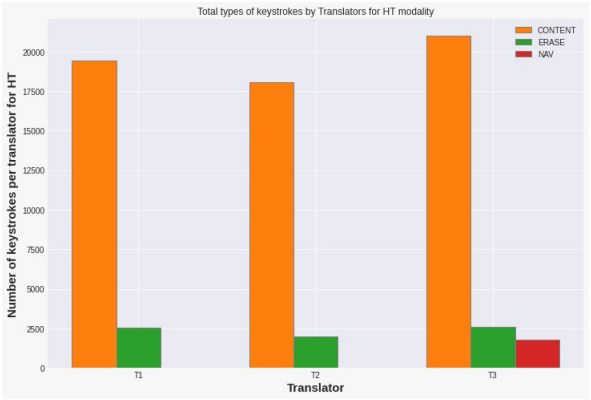

Figure 5: Total count of content, navigation and erase keystrokes for each subject for *ht* modality

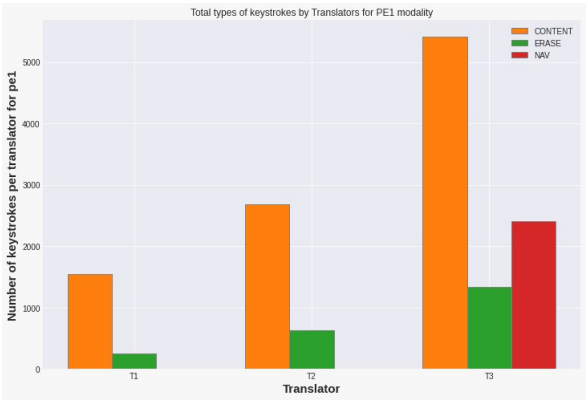

Figure 6: Total count of content, navigation and erase keystrokes for each subject for *pe1* modality

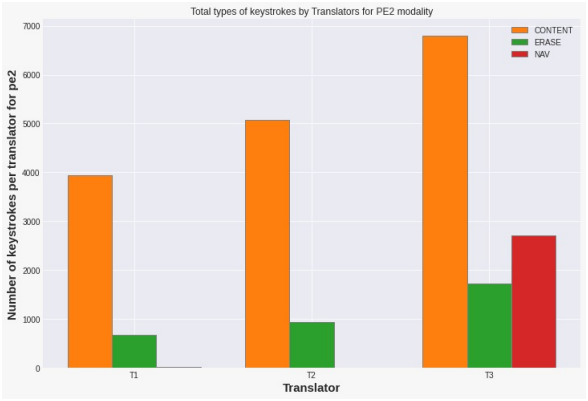

Figure 7: Total count of content, navigation and erase keystrokes for each subject for *pe2* modality

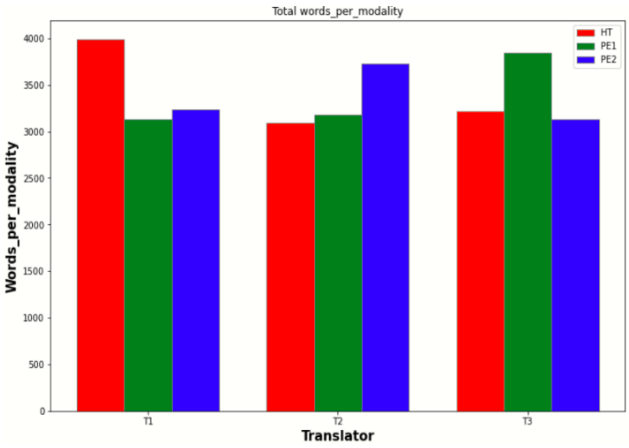

Figure 8: Total count of words for each subject per modality

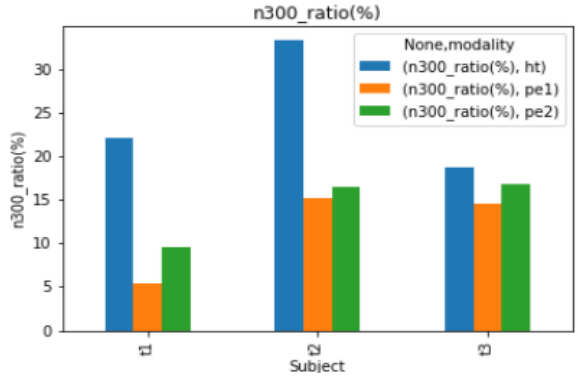

Figure 9: Ratio of the total pausing time (>300ms) with respect to the total editing time

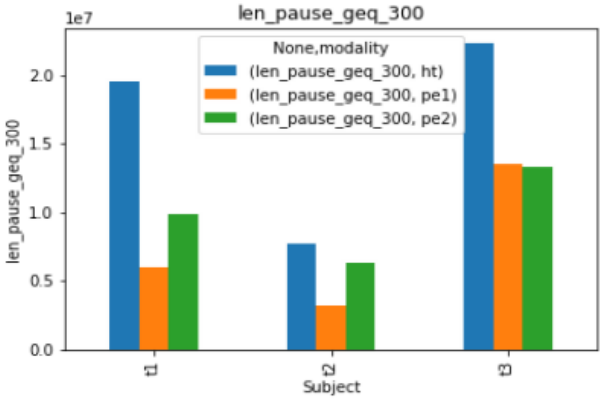

Figure 12: Total length of pauses (>300ms) taken by each subject per modality

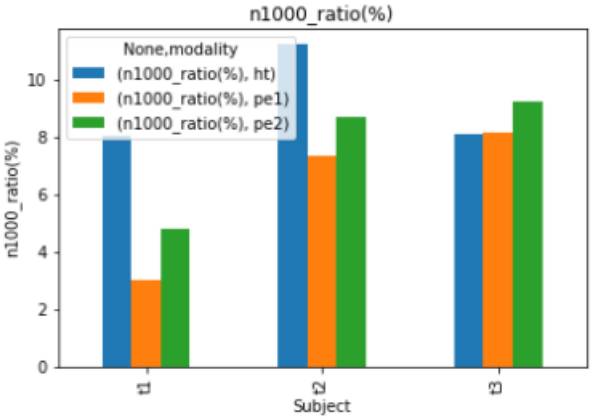

Figure 10: Ratio of the total pausing time (>1000ms) with respect to the total editing time

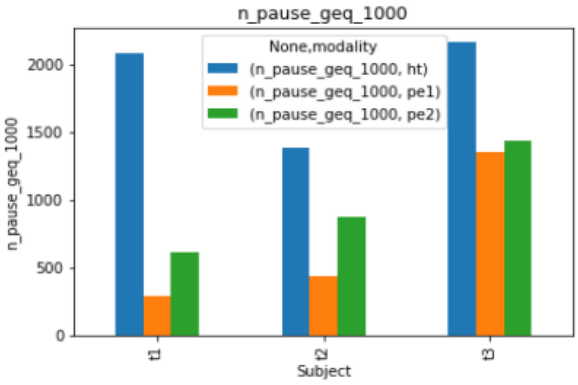

Figure 13: Total number of pauses (>1000ms) taken by each subject per modality

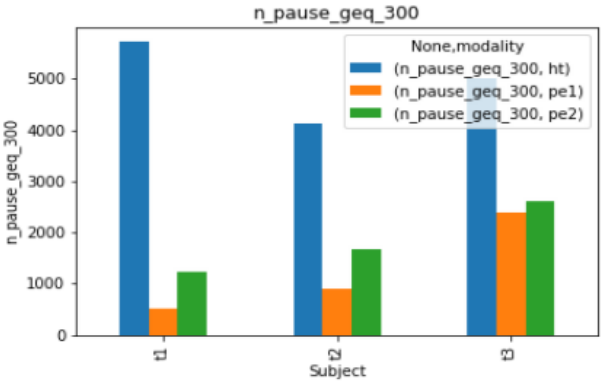

Figure 11: Total number of pauses (>300ms) taken by each subject per modality

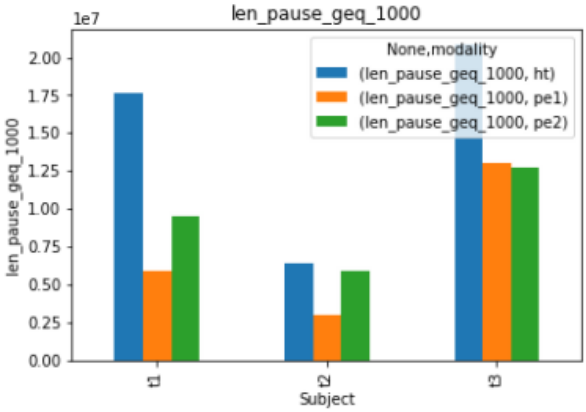

Figure 14: Total length of pauses (>1000ms) taken by each subject per modality

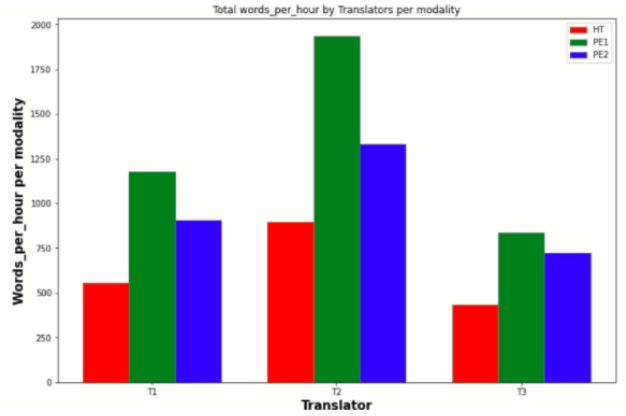

Figure 15: Total words per hour for each subject per modality

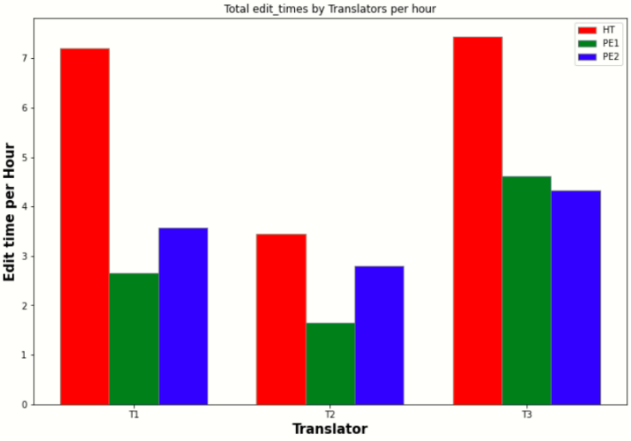

Figure 16: Total edit time per hour for each subject per modality