Stylistic Changes for Temporal Text Classification

Sanja Štajner[1] and Marcos Zampieri[2]

[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
[2]Romance Philology Department, University of Cologne, Germany
sanjastajner@wlv.ac.uk
mzampier@uni-koeln.de

Abstract. This paper investigates stylistic changes in a set of Portuguese historical texts ranging from the 17th to the early 20th century and presents a supervised method to classify them per century. Four stylistic features – average sentence length (ASL), average word length (AWL), lexical density (LD), and lexical richness (LR) – were automatically extracted for each sub-corpus. The initial analysis of diachronic changes in these four features revealed that the texts written in the 17th and 18th centuries have similar AWL, LD and LR, which differ significantly from those in the texts written in the 19th and 20th centuries. This information was later used in automatic classification of texts per century, leading to an F-Measure of 0.92.

Keywords: text classification, stylistic changes, historical corpora, Portuguese

1 Introduction

It is well known that language changes over time. These changes occur in all aspects of language: phonetics, lexicon, grammar, and discourse, as well as in its style. While reading a text dating from a previous century, the reader can often spot that the text contains features that are not common to contemporary language, even if not being aware of its publication date. As it can be seen in [1], studies on lexical and syntactic change are abundant for most languages. The interest of philologists and historical linguists in tracking language change is long-standing, and it exists prior to the development of the first electronic corpora, which are the fundamental resource for current studies in language change.

To the best of our knowledge, very little has been said regarding the stylistic changes of texts. Studies on lexical richness, density and other stylistic aspects of historical texts have mostly been neglected. This is mainly due to the difficulty in quantifying this information before the development of electronic corpora and reliable NLP tools. Only recently have a couple of experiments applied NLP techniques to quantify changes in diachronic corpora [2, 3].

In this paper we investigate stylistic changes of historical texts and use this information to train machine learning algorithms to classify texts automatically. Historical manuscripts are sometimes unidentified regarding its geographical source and/or date of publication, and classification methods can be trained to estimate this information. The methods presented here were applied to a Portuguese historical corpus [4], but they

can be replicated to any language. This study is of interest to researchers in text classification and NLP in general, and historical linguists as well as scholars in the digital humanities who deal with historical manuscripts.

2 Related Work

The vast majority of corpus-based studies on language change focus on grammatical changes (e.g. Leech et al. [5] for English and Galves et al. [6] for Portuguese). A number of English diachronic corpora are available for this kind of study which makes it possible for scholars to use NLP and quantitative methods to examine language change. For Portuguese, only a few resources are available, including: Tycho Brahe [7], and Colonia [4]. On stylistic diachronic changes in 20th-century English language, Štajner and Mitkov [3] report significant changes in several features. Among them – the most relevant for this study – is a significant increase in lexical density and lexical richness in 20th-century British and American English general prose.

Regarding temporal text classification, a couple of studies are worth mentioning. Dalli and Wilks [8] present a computational model to date texts from a time span of nine years. The method is aided by lexical items which increase their frequency at some point of time (e.g. *Bin Laden* in September, 2001 or *World Cup* in June, 2010). The experiments described by Abe and Tsumoto [9] work under a similar assumption. The authors proposed the use of similarity measures to categorise texts based on keywords that are calculated by indexes such as the popular tf-idf. The method obtains document clusters based on temporal differences in the usages of terms.

Mohkov [10] presented one of the systems that participate in the DEFT2010[1] shared task. In this shared task, systems aimed to classify short French journalistic texts of up to 300 words not only with respect to their geographical location, but also regarding the decade in which they were published. Trieschnigg et al. [11] describe a classification experiment using the Dutch Folktale Database. This database includes texts from different dialects and varieties of Dutch, but also historical texts written in middle and 17th-century Dutch. Researchers report a micro average F-measure of 0.799 with the highest F-measure reaching 0.987 for one of the classes.

To the best of our knowledge, the idea of using stylistic features for temporal text classification is new to Portuguese and not substantially explored to most languages. Most studies use lexical and orthographic features to identify the date of publication of a text.

3 Methods

The study consists of two main parts: (1) quantitative analysis of four stylistic features automatically extracted from the corpus; and (2) five text classification experiments.

[1] http://www.groupes.polymtl.ca/taln2010/deft.php

3.1 Corpus

We used the aforementioned Colonia[2] [4], a diachronic collection of historical Portuguese containing texts ranging from the 16th to the early 20th century. The corpus is annotated with lemma and part-of-speech (POS) information, using TreeTagger [12], which is regarded to achieve performance of over 95% accuracy using coarse-grained tags. According to the authors, spelling variation was not systematically normalised, but they acknowledge that some texts presented edited orthography prior to their compilation. At its compilation stage, authors addressed solely the question of unknown lemmas caused by non-standard spelling.

The original Colonia corpus contains 100 texts spanning from 16th to 20th century, balanced between European and Brazilian Portuguese (it contains 52 Brazilian texts and 48 European texts). The time span covered in our experiments comprises the period from 17th to the 20th century and a total of 87 texts. As to the size of the articles, the original corpus contains complete manuscripts of up to 90,000 tokens each. For our experiments, we decided to work with samples of up to 2,000 tokens per text, which were retrieved automatically, starting from a random point in the text (Table 1).

Table 1. Corpora

Century	Texts	Sentences	Tokens
17th	18	1,667	31,635
18th	14	2,566	23,175
19th	38	5,217	63,950
20th	17	2,602	28,569
Total	87	12,052	147,329

We decided to use this sample size in order to obtain results which could be compared with a similar study in English language [3] based on the 'Brown family' of corpora (which also has approx. 2000 tokens per text).

3.2 Experimental Settings

Four stylistic features – average sentence length (ASL), average word length (AWL), lexical density (LD), and lexical richness (LR) – were automatically extracted from the corpora (Table 2). Based on the initial analysis of the distribution of these four features across the four centuries (17th–20th), we decided to conduct five text classification experiments:

1. Classification across all four centuries (17th–20th);
2. Classification between (17th–18th) and (19th–20th) centuries;
3. Classification between the 17th and 18th centuries;
4. Classification between the 18th and 19th centuries;
5. Classification between the 19th and 20th centuries.

Table 2. Features

Feature	Code	Formula
Average sentence length	ASL	ASL = words/sentences
Average word length	AWL	AWL = characters/words
Lexical density	LD	LD = (unique tokens)/tokens
Lexical richness	LR	LR = (unique lemmas)/tokens

All classification experiments were conducted in Weka[3] Experimenter [13], employing four different classification algorithms – Naive Bayes [14]; SMO (Weka implementation of Support Vector Machines) [15, 16] with normalisation and using poly kernels; JRip [17], and J48 (Weka implementation of C4.5) [18] – in 5-fold cross-validation setup with 10 repetitions. In all experiments, we considered the majority class as the baseline.

4 Results and Discussion

The averaged values for each of the four investigated features (ASL, AWL, LD, and LR) in each of the sub-corpora, together with their standard deviations, are presented in Table 3. Statistically significant differences between adjacent centuries are presented in bold. The difference in ASL between the 18th and 19th centuries was reported as statistically significant at a 0.05 level of significance, while all other statistically significant differences were significant at a 0.001 level of significance. Statistical significance was calculated using the two-independent samples t-test in SPSS (in cases where both compared sets followed approximately normal distribution) and using the two-sample Kolmogorov-Smirnov test (in cases where at least one of the sets did not follow approximately normal distribution).

Table 3. Statistics of the corpora (Key: ASL = average sentence length (in words); AWL = average word length (in characters); LD = lexical density; LR = lexical richness)

Century	ASL	AWL	LD	LR
17th	**20.53 ± 6.29**	4.48 ± 0.16	0.38 ± 0.04	0.14 ± 0.02
18th	**11.73 ± 6.42**	**4.52 ± 0.16**	**0.39 ± 0.03**	**0.15 ± 0.02**
19th	**13.73 ± 5.55**	**4.80 ± 0.18**	**0.46 ± 0.03**	**0.19 ± 0.02**
20th	12.79 ± 6.24	4.89 ± 0.32	0.47 ± 0.04	0.18 ± 0.02

The skewness and the existence of outliers can be observed from the box-plots presented in Figure 4. The height of the rectangle indicates the spread of the values for the variable, the horizontal line inside the rectangle indicates the mean, while the "whiskers"

[2] http://corporavm.uni-koeln.de/colonia/
[3] http://www.cs.waikato.ac.nz/ml/weka/

outside the rectangle indicate the smallest and largest observations which are not outliers. Outliers are presented as numbered cases beyond the whiskers. If the rectangle is not equally distributed on both sides of the mean line, then the data is skewed (not normal).

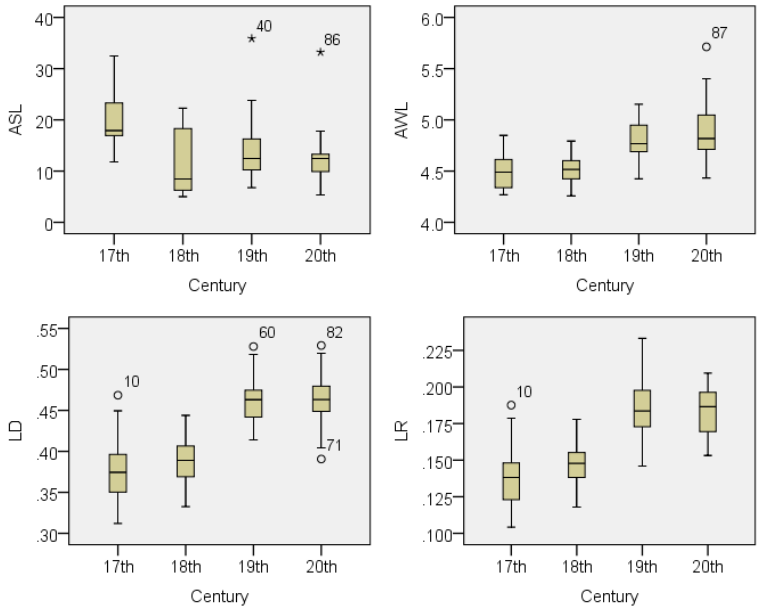

Fig. 1. Distribution of features across the corpora

The results presented in Table 3 and Figure 4 indicate that the average sentence length (ASL) was significantly higher in the 17th than in the 18th century, but then significantly lower in the 19th than in the 18th century. More interestingly, it revealed that the texts written in the 17th and 18th centuries have similar AWL, LD and LR, which were significantly lower than those in the texts written in the 19th and 20th centuries.

These results motivated us to conduct the second text classification experiment (where the texts from the 17th and 18th centuries were grouped together in one class, and those from the 19th and 20th centuries in the other class), in addition to the first classification experiment across all four centuries (17th–20th) and the other three classification experiments between each pair of adjacent centuries (17th and 18th, 18th and 19th, and 19th and 20th).

The results of all classification experiments are presented in Table 4. Columns 'NB', 'SMO', 'JRip', and 'J48' contain weighted average F-measures of the four classification algorithms (Section 3.2), while the column 'baseline' contains the classification accuracy if for each text we select the majority class. Figure 2 contains the rules of the JRip classifier which were used in each of the five experiments.

Table 4. Classification results

Exp.	Classes	NB	SMO	JRip	J48	Baseline
(1)	17th, 18th, 19th, 20th	0.59	0.54	0.52	0.56	0.44
(2)	17th+18th, 19th+20th	0.92	0.92	0.87	0.87	0.63
(3)	17th, 18th	0.64	0.67	0.63	0.73	0.56
(4)	18th, 19th	0.91	0.86	0.88	0.86	0.73
(5)	19th, 20th	0.59	0.57	0.57	0.55	0.69

Experiment I – Classification between the 17th, 18th, 19th, and the 20th century texts:

```
(LD <= 0.421742) and (ASL <= 8.787234) => text=18th (9.0/1.0)
(LD <= 0.407346) and (AWL <= 4.675958) => text=17th (16.0/0.0)
 => text=19th (62.0/24.0)
```

Experiment II – Classification between the 17th+18th and the 19th+20th century texts:

```
(LD <= 0.421742) => text=17th (34.0/5.0)
 => text=19th (53.0/3.0)
```

Experiment III – Classification between the 17th and the 18th century texts:

```
(ASL <= 8.787234) => century=18th (8.0/0.0)
(AWL >= 4.548444) and (AWL <= 4.601594) => century=18th (3.0/0.0)
(AWL >= 4.6875) => century=18th (4.0/1.0)
 => century=17th (17.0/0.0)
```

Experiment IV – Classification between the 18th and the 19th century texts:

```
(LD <= 0.406519) => century=18th (11.0/0.0)
 => century=19th (41.0/3.0)
```

Experiment V – Classification between the 19th and the 20th century texts:

```
(AWL >= 5.186118) => century=20th (3.0/0.0)
 => century=19th (52.0/14.0)
```

Fig. 2. JRip rules for the classification experiments

From the results presented in Table 4, it can be noted that classification accuracies were significantly higher in the second than in the first experiment for all four algorithms, achieving the weighted average F-measure up to 0.92. This is not a surprise given that initial analysis revealed a statistically significant difference in all four features (ASL, AWL, LD, and LR) between the 18th- and the 19th-century texts (Table 3), and the classification between the texts from the 18th and the 19th centuries (experiment 4) achieved almost equally good results. The results of the first and the third experiment, although being significantly lower than those of the second and the fourth experiments, still outperformed the baseline. The results of the classification of texts between 19th and 20th century, however, did not even reach the performance of the baseline. One

possible explanation for this difficulty in classifying texts from these two centuries is that the 20th century class contains only texts published in the first half of the century. The newest text was published in 1948. The style of the texts are therefore very similar to those published in the end of the previous century and this has direct impact on the classifiers' performance.

The difference in the results achieved in the first and the third experiments, and those achieved in the fifth experiment, could also be explained by the fact that the initial analysis of the corpora revealed that there was a significant difference in one of the features (ASL) between the texts from the 17th and the 18th centuries, and there was no significant differences in any of the four investigated features between the texts from the 19th and the 20th centuries. The presented results (Table 3 and Table 4) thus indicate a high correlation between the classification accuracy and the number of features reported to be significantly different between two classes.

5 Conclusions and Future Work

This study was, to the best of our knowledge, a first attempt of comparing the style of historical Portuguese texts in a purely automatic manner. The results indicated similarities between texts from the 17th and 18th as well as the 19th and 20th centuries, and a great dissimilarity between the 18th- and the 19th-century texts. It was also observed that the lexical density (LD) and lexical richness (LR) were substantially higher in the 19th- and 20th-century texts than in the 17th- and 18th- century texts.

As a practical application of our initial analysis of the corpora, we carried out five automatic classification experiments. The first setting containing four classes (one class for each century) achieved a modest 0.59 F-measure which outperformed the baseline. The second setting, binary classification (17th and 18th centuries; and 19th and 20th centuries grouped together), achieved a 0.92 F-measure, thus reflecting the already reported significant differences in all four features (ASL, AWL, LD, and LR) between the texts from the 18th and the 19th century. The lowest classification performances were reported for the classification between the texts from the 19th and the 20th century, again reflecting the fact that the initial analysis of the corpora did not report any significant differences in any of the four investigated features between those two sets of texts.

We continue to experiment with historical texts in different directions. As Portuguese is a pluricentric language, it would be interesting to investigate whether there are significant stylistic differences between these two varieties (both synchronic and diachronic). Previous studies [19] suggest that classification methods are able to distinguish Brazilian and European current texts with 99.8% accuracy when using lexical and orthographic features. It would be worth exploring whether a similar classification accuracy could be achieved by using some language-independent features, thus enabling the use of the same methodology for other languages with their regional varieties.

References

1. Joseph, B., Janda, R.: The Handbook of Historical Linguistics. Blackwell Publishing (2003)
2. Smith, J., Kelly, C.: Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. Computers and the Humanities **36** (2002) 411–430
3. Štajner, S., Mitkov, R.: Diachronic stylistic changes in british and american varieties of 20th century written english language. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria (September 2011) 78–85
4. Zampieri, M., Becker, M.: Colonia: Corpus of historical portuguese. ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research **5** (2013)
5. Leech, G., Hundt, M., Mair, C., Smith, N.: Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press (2009)
6. Galves, C., Sandalo, F.: Clitic-placement in modern and classical European Portuguese. MIT Working Papers in Linguistics **47** (2004) 115–128
7. Britto, H., Finger, M., Galves, C.: Computational and linguistic aspects of the Tycho Brahe parsed corpus of historical portuguese. In: Proceedings of the First Freiburg Workshop on Romance Corpus Linguistics, Freiburg, Germany (2000)
8. Dalli, A., Wilks, Y.: Automatic dating of documents and temporal text classification. In: Proceedings of the Workshop on Annotating and Reasoning about Time and Events, Sidney, Australia (2006) 17–22
9. Abe, H., Tsumoto, S.: Text categorization with considering temporal patterns of term usages. In: Proceedings of ICDM Workshops, IEEE (2010) 800–807
10. Mokhov, S.: A marf approach to deft2010. In: Proceedings of TALN2010, Montreal, Canada (2010)
11. Trieschnigg, D., Hiemstra, D., Theune, M., de Jong, F., Meder, T.: An exploration of language identification techniques for the dutch folktale database. In: Proceedings of LREC2012. (2012)
12. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK (1994)
13. Witten, I., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers (2005)
14. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. (1995) 338–345
15. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation **13**(3) (2001) 637–649
16. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schoelkopf, B., B.C., Smola, A., eds.: Advances in Kernel Methods Support Vector Learning. (1998)
17. Cohen, W.: Fast Effective Rule Induction. In: Proceedings of the Twelfth International Conference on Machine Learning. (1995) 115–123
18. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)
19. Zampieri, M., Gebre, B.G.: Automatic identification of language varieties: The case of Portuguese. In: Proceedings of KONVENS2012, Vienna, Austria (2012) 233–237