TEXT MINING ANALYSIS OF TRANSLATION, SOCIAL COMMUNICATION AND LITERARY WRITING FOR TURKISH

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Sevil Çalışkan

December 2020

TEXT MINING ANALYSIS OF TRANSLATION, SOCIAL COMMUNICATION AND LITERARY WRITING FOR TURKISH

By Sevil Çalışkan

December 2020

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Fazlı Can(Advisor)

Özgür Ulusoy

Fabio Crestani

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

TEXT MINING ANALYSIS OF TRANSLATION, SOCIAL COMMUNICATION AND LITERARY WRITING FOR TURKISH

Sevil Çalışkan
M.S. in Computer Engineering
Advisor: Fazlı Can
December 2020

Text mining is an important research area considering the increase in text generation and the need for analysis. Text mining in Turkish is still not a well-invested research area, compared to the other languages. In this thesis, we analyze different types of Turkish text from different points of views, having an overall review on text mining in Turkish at the end. First, we analyze the translation quality of a Turkish novel, *My Names is Red* novel, to English, French, and Spanish with the features generated for each chapter. With the proposed method, translation loyalties to the original text can be quantified without any parallel comparisons. Then, we analyze the Turkish spoken texts of 98 people in different age groups in terms of gender and age attributes of the speakers. We also analyze the difference between written and spoken texts in Turkish. Results show that it is possible to predict the attributes of the speaker from the spoken text and written and spoken texts are significantly different in terms of stylometric measures. Later on, we make an assessment on cross-lingual transferring performances of multilingual networks from English to Turkish. We see that transferring is possible; however zero-shot cross-lingual transferring still has its way to be competitive with monolingual networks for Turkish. Lastly, we conduct a time-based stylometric analysis of Ahmet Hamdi Tanpınar's works. We see that Ahmet Hamdi Tanpınar shows some differences compared to his contemporaries.

Keywords: text mining, stylometric analysis, spoken text analysis, discourse analysis, cross-lingual learning, transfer learning, multi-lingual data.

ÖZET

METİN MADENCİLİĞİ İLE TURKÇEDE ÇEVİRİ, SOSYAL İLETİŞİM VE EDEBİ YAZI ANALİZİ

Sevil Çalışkan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Fazlı Can
Aralık 2020

Metin madenciliği, üretilen metin miktarının düzenli olarak artması ile birlikte geniş ve önemli bir araştırma konusu haline gelmiştir. Türkçede metin madenciliği araştırmaları ise, diğer diller ile yapılan çalışmalar ile karşılaştırıldığında, hâlâ yeterli değildir. Bu tez çalışmasında, farklı türlerdeki Türkçe metinleri, farklı açılardan ele alınarak çeşitli metin madenciliği yöntemleri ile analiz edilmişlerdir. İlk çalışmada, *Benim Adım Kırmızı* romanının metni ve bu romanın İngilizce, Fransızca ve İspanyolca çevirileri kullanılarak çevirilerin aslına sadakatinin değerlendirilmesi yapılmıştır. Önerilen metod ile çevirilerin aslına bağlılığı, doğrudan cümle çevirilerine ihtiyaç duymadan değerlenlendirilebilir. Tezin ikinci kısımında, 98 kişinin konuşma metinleri, bu kişilerin cinsiyet ve yaş özellikleri yönünden analiz edilmiştir. Sonuçlar, kişilerin konuşma metinlerinden cinsiyet ve yaş tahmini yapılabileceğini doğrular niteliktedir. Daha sonra çok-dilli yapay ağların çapraz dil aktarımı performanslarını, metin sınıflandırma için ölçen bir çalışma yapılmıştır. Deney sonuçları, başka dillerden Türkçeye aktarım olabileceğini gösteren niteliktedir fakat hâlâ eğitim için hazırlanmış veri olmadan, doğrudan başka dillerden taşıma yeterince başarılı değildir. Son olarak Ahmet Hamdi Tanpınar'ın eserlerinin zamansal olarak üslup analizi yapılmıştır. Analiz sonuçları, Ahmet Hamdi Tanpınar'ın üslup değişimi açısından, çağdaşlarından farklılık gösterdiğini ortaya koymuştur.

Anahtar sözcükler: metin madenciliği, konuşma metini analizi, aktarımlı öğrenme, çok-dilli veri, çapraz dil öğrenimi.

Acknowledgement

Firstly, I would like to thank my advisor Prof. Fazlı Can for his guidance, support and understanding throughout my master's studies. It was a great honor to work with him. He was more than just an advisor, and it was always nice to have a lovely chat with him about any intellectual topic. Besides my advisor, I would like to thank the rest of my thesis committee for their time.

I also thank Prof. Hasan Akbulut, Prof. S. Ruken Öztürk, Prof. Emine Uçar İlbuğa and Assist. Prof. Mert Gürer for the dataset of the work in Chapter 3. Their study was supported by the TÜBİTAK project 115K269 and the Istanbul University Scientific Research Projects Coordination Unit project 25080. The study in Chapter 3 was done with the data produced from the TÜBİTAK project. I am also grateful to the other researchers of the TÜBİTAK project, for their work as interviewers; and a number of students for recording and transcribing the interviews.

I would like to thank my dearest friend Ece, for her constant support and love from the first day we met to today, including my master's. I also thank Aysel, Negin, Erdem, Furkan, Baturay, Hakan for listening me whining about any problem I had all over this three years and of course for their valuable friendship. I would also like to thank all my friends in Bilkent for the good memories and good coffee.

Lastly, I would like to thank my parents and siblings, whose love is always with me, for supporting and encouraging me throughout this experience. I also thank my niece and nephew for just existing and making days brighter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis introduces new text mining studies conducted with Turkish. Analyses use stylometric and machine learning methods, on literary writings, spoken texts and multi-lingual data.

1.1 Context of the Study

Stylometry is used for authorship attribution, author verification, author profiling, date attribution, and other related problems. Studies conducted on Turkish texts generally aim to identify the author of a new text by using newspaper columns of different authors. [3]. Moreover, by using the texts of the columnists who wrote in various areas, a new text in a particular area is aimed to be attributed [4]. For author verification purposes, studies conducted looking for evidence that two Turkish texts are written by the same author [5] It is known that the use of language may vary according to age or gender of authors [6]. Author profiling aims to narrow the search space by determining the author's gender, age, and other demographic characteristics. Date attribution tries to find the date on which a text is written. It can also be used to examine language and writer style

change with time. [7, 8]. Stylometric analysis involves the following steps: Preprocessing, feature extraction, classification, and performance assessment. The analysis is possible by converting a text into some numeric features or attributes. Various features are used by researchers, and they are grouped in different ways. Most common groupings can be found as lexical features, character-based features and syntactic features. Many other domain-specific features and their groupings are proposed [9]. When selecting attributes, care should be taken to avoid loss of information. Classification approaches are quite diverse. Most papers use machine learning-based, distance-based, probabilistic and statistical methods. Many distance-based, and probabilistic approaches can be considered as machine learning methods as well. After selecting the classifier, results should be compared with proper baselines.

Discourse is defined as naturally occurring language [10] and in the last few decades, with the proliferation of social media platforms, it is recorded online. As more analysis tools become available, discourse is more often analyzed quantitatively by researchers. Large amounts of discourse data are analyzed to extract the demographic information of their owners [6, 11, 12], or for use in tasks like opinion mining [13]. Even though written discourse on social media is closer to everyday speaking language, it can still vary. Given that spoken discourse is now captured and stored easily by many electronic devices, the application of quantitative analysis methods on spoken language data may reveal information regarding the owners of the discourse, or the discourse itself and can be promising for many areas including social sciences, artificial intelligence studies, marketing, and customer services.

Given the significant improvements in NLP domain in the last years, researchers now use multi-lingual data together. Many studies are invested in the topics of multi-lingual and zero-shot learning with promising results [14–16]. Since many language suffer from the lack of data for specific tasks, such studies encourage researchers to use methods that can benefit data of multiple languages. multi-lingual data is used for many tasks like natural language inference, machine translation, classification and etc [17]. In this study, we focus on the problem of micro-text classification with limited or no data in a given language. Using data

from another language for training purposes, is a common and applicable solution, where data collection is not possible. Most proposed method is to translate the data of languages to a mediator language or one to another [18, 19]. However, it may cause information loss in data and is not practical since translating adds an extra step to classification. Another method is mapping word embedding of to languages to the same space, which allow transferring from one language to another [20–22]. With the lately proposed, large and multi-lingual networks, zero-shot cross-lingual cross-lingual transfer can be used for similar purposes [14], Zero-shot transfer simply means training a model in a source language, and transferring to a target language. Models like BERT and XLM-R are pre-trained with data of many languages together and allow zero-shot learning with impressive performances on multi-lingual cases [23, 24].

1.2 Description of Thesis Content and Contributions

In this thesis, we apply 3 different analyses on Turkish data from different domains. In the second chapter, we aimed to evaluate the loyalty of translations for Orhan Pamuk's novel *My Name is Red* to English, French, and Spanish. For this purpose, different sections in the novel are thought as to be from the mouth of different fictional characters. The hypothesis was the styles of characters will be preserved to some point in the translations if the style has not changed consciously by translator. To test this hypothesis; firstly, attribute vectors of characters were created for original text and target language translations, and the distances between original and translation texts were calculated by these vectors. Contribution of this chapter is, it proposes a new method to evaluate the loyalty of translations with stylometric measures.

In the third chapter, we study the spoken memoirs of a group of old-time moviegoers of different age groups from Turkey. Cinema-going experiences and memories of an audience are seen as a cultural heritage, where cultural heritage

is known as an expression of the ways of living developed by a community and passed on from generation to generation, including customs, practices, places, objects, artistic expressions, and values [25]. The study that provides the text employed in this article fundamentally considers such cultural aspects [1]. In the work presented, the responses of old-time moviegoers are investigated from the viewpoint of discourse analysis with the expectation that various attributes of a participant are reflected by their everyday speaking language. The contributions of this chapter can be highlighted both in terms of its results and the possibilities brought by its newly introduced dataset. We provide the first quantitative analysis of discourse based on a simultaneous, naturally spoken unedited, text for the Turkish language. All participants, answered the same set of questions spontaneously: The context of the talks is the same, which eliminates possible language use variations that may be introduced by topic variations. The dataset we provide in this study opens the door to many research possibilities, since it can be annotated in different ways for various NLP-related studies.

In the fourth chapter, we use multiple Turkish micro-blog datasets for classification. We use English micro-blog datasets from the same and different domains, to investigate zero-shot transfer performance of multi-lingual networks, BERT and XLM-R, on Turkish data. Our contribution with this study is, to assess the zero-shot cross-lingual abilities of multilingual networks and compare performances of the multilingual networks with newly released monolingual networks for Turkish in the classification tasks.

In the fifth chapter, we analyze the different types of works of Ahmet Hamdi Tanpınar. We use six style measures: "sentence length in terms of words", "most frequent words", "word length of type", "word length of token", "syllable count of type" and "syllable count of token". We examine the change in sentence length, word length and most frequent words with time.

In the final chapter, we conclude the thesis with a summary of each chapter and future research possibilities for them.

Chapter 2

Quantification of Loyalty for Translations of *My Name is Red*

2.1 Introduction

In this chapter, we introduce stylometry, an important problem of digital humanities. We present a novel study on quantification of translation loyalty by using a work of Orhan Pamuk: *My Name is Red*. An earlier version of this chapter is published in *Turkish Librarianship* in December 2018 [26]. The tables and the figures of this chapter are taken from ref with the permission of the publisher.

2.1.1 Stylometric Analysis

Stylometry is used for authorship attribution, author verification, author profiling, date attribution, and other related problems. Stylometric analysis involves the following steps: Pre-processing, feature extraction, classification, and performance assessment. The analysis is possible by converting a text into some numeric features or attributes. Converting all properties of a text into numeric

attributes can increase the computation time; and reduce the classification accuracy. The use of right features usually decreases the number of attributes and computation time. Therefore, a text is subject to some pre-processing prior to obtaining the attributes.

Various features are used by researchers, and they are grouped in different ways. Most common groupings can be found as lexical features, character-based features and syntactic features. Many other domain-specific features and their groupings are proposed [9]. When selecting attributes, care should be taken to avoid loss of information. On the other hand, every feature of the text may not always be useful. To evaluate the relevancy, measures like information gain, gain ratio, symmetrical uncertainty, correlation are used [27].

Classification approaches are quite diverse. Many studies use machine learning-based, distance-based, probabilistic and statistical methods. Many distance-based, and probabilistic approaches can be considered as machine learning methods as well. After selecting the classifier, results should be compared with proper baselines.

2.1.2 Background and Work Done

In the translation texts, it is aimed to express the original text in another language without changing the meaning ofit. So, can the style be preserved while preserving the meaning. Studies are available in the literature which examine the translation texts in terms of style loyalty. Can et al., aims to examine the style relation between Shakespeare sonets and the translations to Turkish numerically [28]. Patton et al, tries to determine the unchanging style features of James Joyce's *Dubliners* stories and the translation to Turkish [29]. Baker examines translations of the same text by using stylistic analysis and looks for traces of different translators [30].

In this study, it is aimed to evaluate the loyalty of translations of Orhan Pamuk's novel *My Name is Red* to English, French, and Spanish. For this purpose,

different chapters in the novel are used, where each chapter is from the voice of a different novel character. The hypothesis was the styles of characters will be preserved to some point in the translations if the style has not changed consciously by translator. To test this hypothesis; firstly, attribute vectors of characters were created from original text and target language translations, and the distances between original and translation texts were calculated by these vectors.

2.2 Method

My Name is Red is composed of 59 chapters and each chapter is written in the voice of one of 20 different characters in the novel. The list of characters for original text and translations can be seen in Table 2.1. The order of the characters in the table is the same as the order as they first appeared in the novel. Translations were made from Turkish to English by Erdağ M. Göknar, to French by Gilles Authier and to Spanish by Rafael Carpintero.

The voices of characters are divided into chapters in the novel, allowing us to examine the style of characters separately. In order to evaluate the stylistic similarity between original text and translations in numerical terms, firstly the chapters of the same characters are combined. With this merge, it is aimed to create the widest dataset for characters.

In order to compare the styles of the characters, texts were transformed into numerical forms, in other words attributes were obtained from the texts. In this study, lexical features were used as attributes, which are number of words (number of tokens), number of different words (number of types), average word length in letters for both types and tokens, average sentence length in words, average count of vowels for both token and types, and frequency of the most frequently used words (vowel count is equal to syllable count in Turkish). While count of vowels are calculated, only the letters are considered, and their sounds are ignored. When creating frequency vectors, the most frequently used k words of each character were merged, and repeating words were removed. Since repeating

Table 2.1: Table of characters in the novel *My Name is Red.*

Turkish	**English**	**French**	**Spanish**
Ben Ölüyüm	I am a corpse	Je suis mon cadavre	Estoy muerto
Benim Adım Kara	I am called Black	Mon nom est Le Noir	Me llamo Negro
Ben, Köpek	I am a dog	Moi, le chien	Yo, el perro
Katil Diyecekler Bana	I wıll be called a Murderer	On m'appellera l'Assassin	Me llamarán Asesino
Ben Eniştenizim	I am your beloved Uncle	Je suis votre Oncle	Soy vuestro Tío
Ben, Orhan	I am Orhan	Moi, je m'appelle Orhan	Yo, Orhan
Benim Adım Ester	I am Esther	Mon nom est Esther	Me llamo Ester
Ben, Şeküre	I, Shekure	Moi, Shékuré	Yo, Seküre
Ben Bir Ağacım	I am a tree	Je suis l'arbre	Soy un árbol
Bana Kelebek Derler	I am called "Butterfly"	On m'appelle Papillon	Me llaman Mariposa
Bana Leylek Derler	I am called "Stork"	On m'appelle Cigogne	Me llaman Cigüeña
Bana Zeytin Derler	I am called "Olive"	On m'appelle Olive	Me llaman Aceituna
Ben, Para	I am a gold coin	Moi, l'Argent	Yo, el Dinero
Benim Adım Ölüm	I am Death	Mon nom est la Mort	Me llamo Muerte
Benim Adım Kırmızı	I am Red	Mon nom est Rouge	Me llamo Rojo
Ben, At	I am a horse	Moi, le Cheval	Yo, el caballo
Üstat Osman, Ben	It is I, Master Osman	Moi, Maître Osman	Yo, el Maestro Osman
Ben, Şeytan	I, Satan	Moi, le Diable	Yo, el Diablo
Biz, İki Abdal	We two dervishes	Nous, les deux Errants	Nosotros, dos derviches errantes
Ben, Kadın	I am a woman	Moi, la Femme	Yo, la mujer

words will be different for each language, size of word frequency vectors differs in original text and in translations. Lexical properties and word frequency vectors were then combined with different combinations to form feature vectors for each character. There are 20 attribute vectors matching the number of characters.

We have listed eight lexical features and first seven of them will be referred as lexical features and the last one will be referred as the most frequent words frequencies, for the ease of expression.

Based on the assumption that the styles of the characters will be preserved in the translations where the style is preserved, it is expected that the distances between the attributes of the characters calculated with the original text show the same distribution in translations. In order to test it numerically, the distances between the character vectors for the original text and each translation are calculated as Euclidean distance on the PCA plane. Correlation values between these distances were calculated and the relationship between original text and translations was observed. We refer to this correlation coefficient values as rank consistency-based similarities (RCS). Kendall's Tau Coefficient (Kendall's Rank Correlation Coefficient) was used for correlation calculation. Rank correlation measures the degree of similarity between the two ordered variables and evaluates the significance of this relationship statistically. Kendall's Tau Coefficient was chosen because it can measure the similarity between the two variables without the need for information on the distribution of variables. In this step, distances among the translations and original text can be calculated as well, without reflecting them on a PCA plane. Since the rankings will be the same, results would not differ.

2.3 Experimental Results

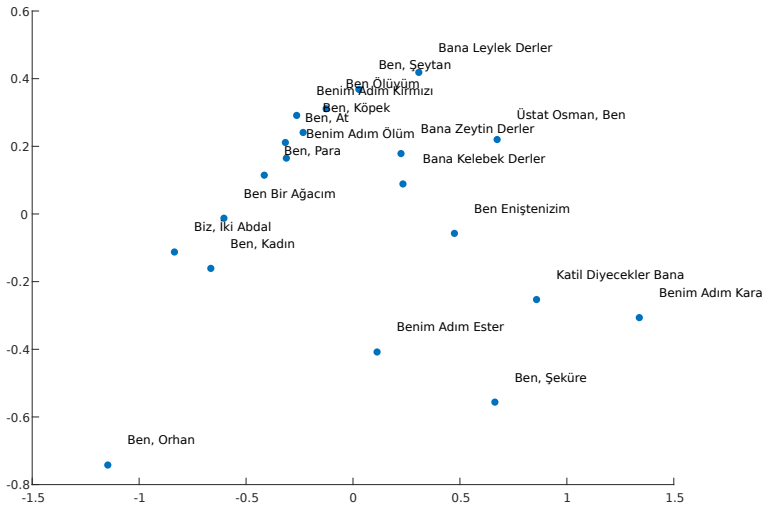
Following the extraction of the attribute vector for each character, the principal component analysis (PCA) was applied to the vectors in order to observe the stylistic similarity of the characters. The principal component analysis is a size

reduction tool that can be used to reduce a large set of variables to a small set containing most of the information in the set.

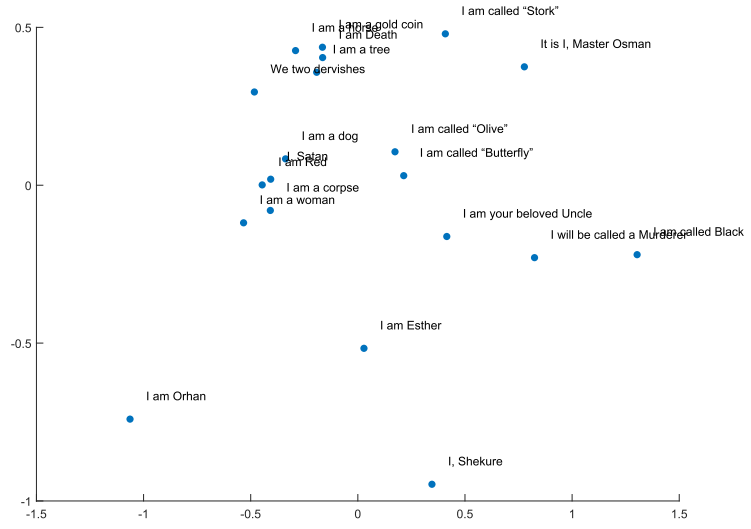
Figure 2.1 shows 4 different scatter plots. It can be observed that the distribution of the stylistic features of the characters are similar for the original text and translations. It is possible to observe in the plots that number of words in the text segments corresponding to novel characters. is quite effective to differentiate them. However, it is not the only strong factor. For example, the features for the number of words of Orhan character and Two Abdals (851, 755) are close to each other and the number of words of Ester and Shekure (8723, 18404) is quite large compared to of Orhan. However, Orhan's distance from Two Abdals, does not seem very different from the distance from Shekure and Ester. For this reason, experiments have been carried out with different properties in order to examine the effect of attributes on the relationship between original text and translations.

The first experiment was made with feature vectors created using all lexical features but not frequency vectors. In the experiment, the style relationship between translations was aimed to be observed directly. The most commonly used words can be an effective determinant for the characters, so they can hide the effects of other stylistic features. That is why, they are not included in the attribute vector in this experiment. Correlation values among the original text and the translations can be seen in Table 2.2 As the p-values for all the correlation coefficients calculated in the article were smaller than 0.001, it was not required to make a table of them. The p-value indicates the probability that the calculated correlation values are coincidental. In this case, a small p-value indicates that this is unlikely to be a coincidence.

It is expected that the number of words and the number of different words will increase the correlation with the assumption that they will change in similarly for characters in the original text and translations. In Table 2.2.b, the results were shown for the experiment conducted with the feature vectors from which the number of words and the number of different words were removed. All correlation values decreased compared to Table 2.2.a. The results confirm the expectation.

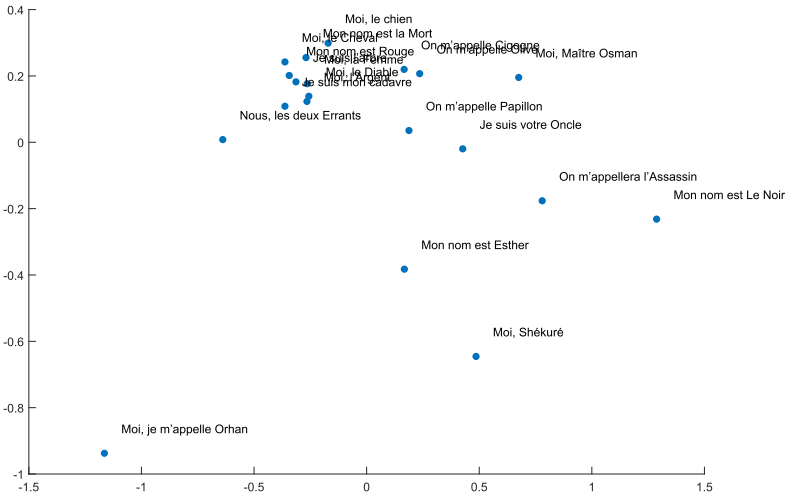


(a)

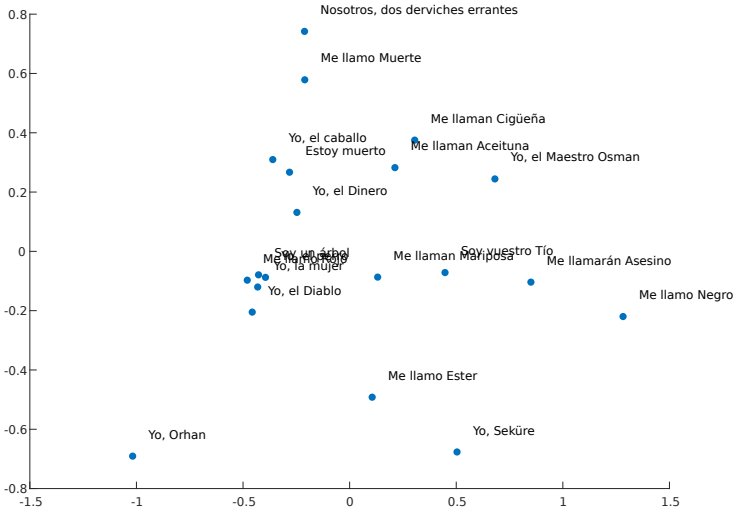


(b)

Figure 2.1: PCA diagram created with lexical features for a. Turkish, b. English, c. French and d. Spanish.



(c)



(d)

Figure 2.1: PCA diagram created with lexical features for a. Turkish, b. English, c. French and d. Spanish.

Table 2.2: RCS values calculated with lexical features (a, left), and correlations after words counts and different word counts are removed from lexical features (b, right).

	Turk	Eng	Fren	Span
Turk	1	0.649	0.720	0.634
Eng		1	0.741	0.701
Fren			1	0.714
Span				1

	Turk	Eng	Fren	Span
Turk	1	0.533	0.629	0.513
Eng		1	0.655	0.586
Fren			1	0.621
Span				1

As mentioned, the most commonly used words can be effective in separating characters. In order to confirm this, as mentioned earlier experiments were performed with the most frequent words vectors for different k numbers. In Table 2.3.a results for k= 10, in Table 2.3.b for k= 20, and in Table 2.3.c for k= 30 can be seen. When the results of the experiments are compared with Table 2.2.a, it is seen that lexical features without the frequency of most frequent words give better results for all k values. More interestingly, the most common words vectors give the best coefficient values between the original text and the translations for the k= 20 value. This may be due to a few of the most frequently used words may not capture distinctive words, and when they are too many, distinctive words may appear in other characters as well.

Table 2.3: RCS values calculated with the most frequent words vectors k = 10, k = 20, and k = 30. (a, b, c)

k=10	Turk	Eng	Fren	Span
Turk	1	0.455	0.399	0.241
Eng		1	0.585	0.436
Fren			1	0.391
Span				1

k=20	Turk	Eng	Fren	Span
Turk	1	0.672	0.654	0.680
Eng		1	0.588	0.674
Fren			1	0.640
Span				1

k=30	Turk	Eng	Fren	Span
Turk	1	0.644	0.421	0.568
Eng		1	0.616	0.697
Fren			1	0.673
Span				1

The results of the experiments performed to observe the similarity between the texts as the k increases can be seen in the plots of Figure 2.2. While the similarity of the translations with the original text continues to change as the k values increase, the similarity between the translations reaches a steady state. The fact that English, French and Spanish are from the European branch of the Indo-European family of languages and Turkish is from a different family, the Ural-Altaic language family, may be the reason for these observations.

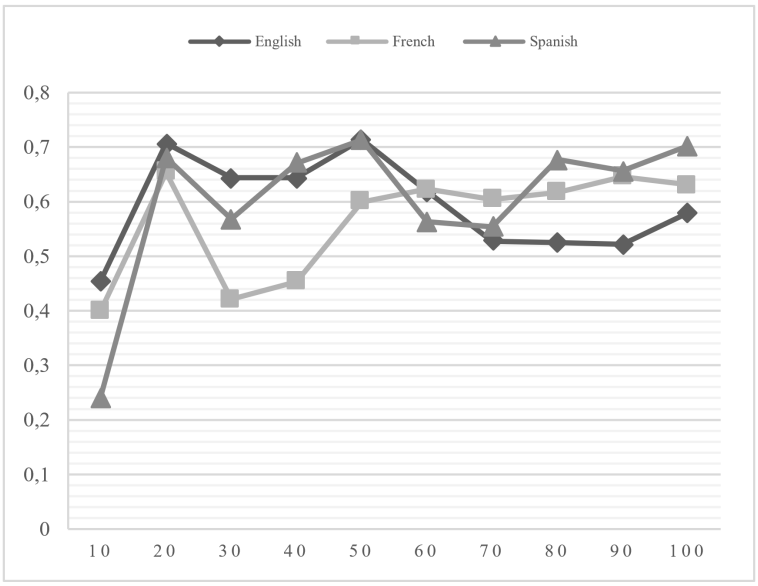
The results of the experiment with the combination of lexical features and the most frequent words vectors can be seen in Table 2.4. Compared to Table 2.2.a and Table 2.3.b, these vectors appear to yield better results when they are combined.

When the coefficients are examined, the similarity of the English, French and Spanish translations is usually higher than the similarity to the original text in Turkish. For example, in Table 2.4, the coefficient between the original Turkish text and the Spanish translation is 0.7069, while the coefficient between the English translation and the Spanish translation is 0.7194. This may be, again, the result of the translation languages being from the same language family. Different numerical analyses were performed to examine the differences between Turkish and these languages.

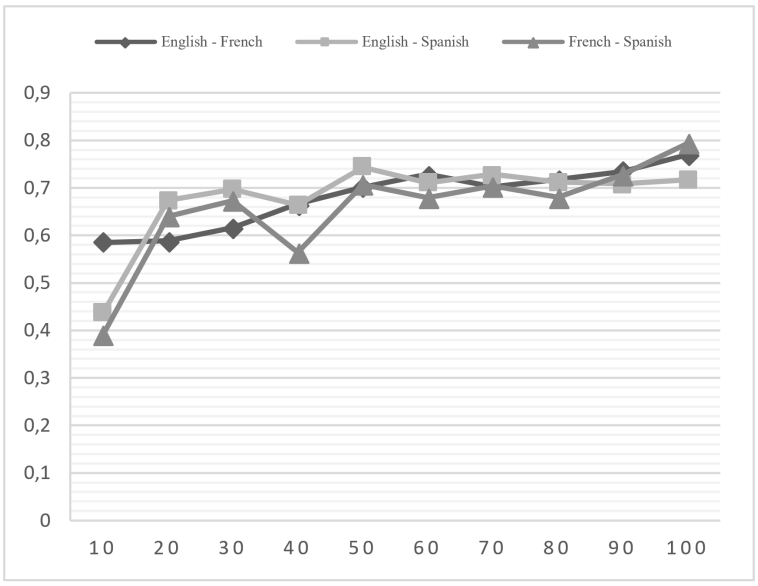
Table 2.4: RCS values calculated with lexical features and the most frequent words frequencies vectors (k = 20).

	Turk	Eng	Fren	Span
Turk	1	0.707	0.675	0.705
Eng		1	0.656	0.719
Fren			1	0.688
Span				1

The first analysis is the calculation of the word count rates for the characters. In Table 2.5, word counts of the texts in the language of the rows are divided by the number of words in the language of the columns and the result is written in the cell where the row and the column intersect. These ratios are calculated for each character and averaged over the characters. According to the results, it



(a)



(b)

Figure 2.2: RCS value graphs for different k values between Turkish (a) and translation (b) texts.

Table 2.5: Averages ratios of (a) number of words and (b) number of different words in terms of the characters.

	Turk	Eng	Fren	Span
Turk	1	0.640	0.603	0.617
Eng		1	0.944	0.966
Fren			1	1.025
Span				1

	Turk	Eng	Fren	Span
Turk	1	1.223	1.067	1.172
Eng		1	0.882	0.967
Fren			1	1.100
Span				1

can be interpreted that more words are used to express the same situations in English, French and Spanish. In Turkish, time or subject is expressed by suffixes, while in the translation languages usually different words are used for them. In short, results can be explained by the Turkish language being an agglutinative language.

The proportions of the different word numbers for the characters were calculated by using the different languages in the same way and then the averages were taken. As can be seen from Table 2.5.b, in contrast to the total number of words, the number of different words used in Turkish is higher than in English, French and Spanish. This situation can also be explained by the fact that Turkish is an agglutinative language and many different words are produced using the same root. After Turkish, French shows more varied vocabulary than other languages.

Table 2.6: Average sentence length ratios.

	Turk	Eng	Fren	Span
Turk	1	0.673	0.653	0.635
Eng		1	0.972	0.945
Fren			1	0.976
Span				1

The average sentence length ratios were calculated as in Table 2.5 with sentence lengths and the results were given in Table 2.6. Examining the table, it can be interpreted that the sentences in Turkish are shorter than the translation languages. On the other hand, sentence lengths in translation languages are not very different. The sentence length ratio for Turkish - English overlaps with the ratio of 0.66 in the study of Patton and Can [29].

2.4 Discussion

In this study, it is aimed to evaluate the style similarity between *My Name is Red* novel and its translations. For this purpose, style of the fictional characters in the novel were examined numerically. With lexical features and most frequent word frequencies, experiments were conducted. In the experiments performed with lexical features and the most frequently used words separately, correlation coefficient of the original text with the translations was close to each other. Best result was seen when those features were merged as a single attribute vector. This is an expected result since different features have the potential of bringing new information about the novel character.

Experiments with the most frequently used words were continued by expanding the number of k, and as the k-value increases, the correlation between the translations remained constant between 0.7 and 0.8. No significant improvement was observed with the increase of k value in the similarity between original Turkish text and translations. It can be explained by the fact that Turkish has a different origin from the translation languages.

When we look at the ratio of the number of words between Turkish texts and translations, it is seen that the number of words used in the translation languages is higher. Since the Turkish language is an agglutinative language, this result is not different from expected. The different word ratios used in the texts indicate that the number of different words used in Turkish is higher than the translations. In addition, translation languages are almost identical to each other in terms of different word numbers. In sentence length ratios, the rates of translation languages are also close to each other. The translation languages being the members of the same language family can also explain these results.

2.5 Conclusion and Future Work

In this chapter, we present a new work on the quantification of loyalty for translations. The style similarity between original text of *My Name is Red* and its translations in English, French and Spanish is analyzed by using a rank consistency-based measure that we introduce in this paper. The experiments with lexical features and most frequently used words statistically significantly confirm the similarity. They also show that the pairwise compatibility of translation languages is higher than that of their compatibility with the Turkish original. The observations are as expected for these target languages that are the members of the same language family.

In future research, experiments with different language families can be performed. Furthermore, similar studies can be done with other works by dividing them into cohesive units, such as short stories or parts of novels containing related themes.

Chapter 3

Quantitative Analysis of Spoken Discourse Using Memoirs of Old-time Moviegoers

This chapter of the thesis is a version of the journal paper under review in *Journal of Quantitative Linguistics* with minor revision [31].

3.1 Introduction

Discourse is defined as naturally occurring language [10] and in the last few decades, with the proliferation of social media platforms, it is recorded online. As more analysis tools become available, discourse is more often analyzed quantitatively by researchers. Large amounts of discourse data are analyzed to extract the demographic information of their owners [6, 11, 12], or for use in tasks like opinion mining [13]. Even though written discourse on social media is closer to everyday speaking language, it can still vary. Given that spoken discourse is now captured and stored easily by many electronic devices, the application of quantitative analysis methods on spoken language data may reveal information

regarding the owners of the discourse, or the discourse itself and can be promising for many areas including social sciences, artificial intelligence studies, marketing, and customer services.

In this chapter, we study the spoken memoirs of a group of old-time moviegoers of different age groups from Turkey. Cinema-going experiences and memories of an audience are seen as a cultural heritage, where cultural heritage is known as an expression of the ways of living developed by a community and passed on from generation to generation, including customs, practices, places, objects, artistic expressions, and values [25]. The study that provides the text employed in this chapter fundamentally considers such cultural aspects [1]. In the work presented, the responses of old-time moviegoers are investigated from the viewpoint of discourse analysis with the expectation that various attributes of a participant are reflected by their everyday speaking language.

We use stylometric and statistical tools as analysis methods. Analysis of the discourse data is guided by various research questions. The first question answers whether the age and gender of the participant can be inferred from their response (spoken text), as in the case of written text. The second question asks if there are differences in language use between younger and older people in terms of vocabulary richness and archaic word usage. Based on the results of previous studies [32] we expect that older generations will use words shorter than new words. In Turkish this length difference is attributed to the fact that the meanings of old words, usually taken from Arabic or Persian, were replaced with words with Turkish roots followed by suffix(es) [33], such as the replacement of "tebessüm" with "gülümseme" meaning "smile". Distinguishing spoken or written text becomes the third research question. Lastly, considering that there are people among the participants who write as well, the question of the ability to distinguish spoken text of a person from their written text arises.

The contributions of this chapter can be highlighted both in terms of its results and the possibilities brought by its newly introduced dataset. We provide the first quantitative analysis of discourse based on a simultaneous, naturally spoken unedited, text for the Turkish language. All participants, answered the same set

of questions spontaneously: The context of the talks is the same, which eliminates possible language use variations that may be introduced by topic variations. The dataset we provide in this chapter opens the door to many research possibilities, since it can be annotated in different ways for various NLP-related studies.

The rest of the chapter is organized as follows: Section 3.2 includes related works on the analysis of spoken discourse. Section 3.3 introduces our dataset. Section 3.4 proposes the research questions of the chapter and sets up our experimental design. Section 3.5 shares the experimental results. Section 3.6 provides future work pointers and concludes the chapter.

3.2 Related Work

Language usage is studied in many different areas including psychology [34], linguistics [35], socio-linguistics [36], and, with the wide use of social media, in computer science [37–40]. However, in this chapter, our concern is the quantitative analysis of spoken text; therefore, we only provide a short review of works in this area.

3.2.1 Differences of spoken and written language

Studies of spoken language is dated to an era where machine learning was not greatly employed. Early works usually studied the differences or relationships between spoken and written language from a linguistic point of view [41–44]. It is shown that differences in spoken and written discourse exist even in the case of context similarity [41]. Factor analysis is greatly used to detect variations in the linguistic dimensions of speech and writing [42]. There are numerous studies on this topic as provided in [43].

3.2.2 Differences in spoken language among different groups

Differences in spoken language use among different groups of people always attract the attention of researchers from different areas and communities as a study topic. The change in language use with the age and gender is widely studied. In one of the studies, the telephone conversations of participants between the ages of 17 to 68 are analyzed to find patterns of age-related change in conversational speech [45]. In a similar study, the relation of age and the frequency of selected nouns is investigated [46]. Language change with age is also surveyed from a socio-linguistic perspective focusing on age and aging as a social attribute [47]. Gender is considered as another socio-linguistic variable and surveys are concentrated on variation studies [48]. Social interaction is showed to differ with gender and it is discussed that language use also changes with the change in the gender of the audience [49]. Differences in speech style and behavioral responses during speaking are studied focusing on gender [50, 51]. Even popular science books are published regarding gender differences in language [52]. Almost all of the studies include statistical analysis for supporting the validity of their claims.

3.2.3 Analysis of the spoken language of politicians

Lately, the language usage of politicians is analyzed widely. The 2016 US election speeches and campaigns of candidates were very popular as a research topic. Savoy examines the styles of two main candidates, Donald Trump and Hillary Clinton, and with various global style markers, observes that Trump's oral style is more direct while Clinton uses longer sentences with a richer vocabulary [53]. Similarly, lexicon-based sentiment analysis is applied to the speeches of both candidates [54]. The authors attribute thematic words for each candidate using word embeddings and their results find that Trump's speeches are more negatively annotated than Clinton's in terms of sentiment. Donald Trump's political discourse is evaluated based on the stylistic features by comparing it to the discourses of

Obama and Clinton [55]. The authors compare thematic concentration, readability levels of debates, and vocabulary richness using ANOVA and a set of non-parametric tests. The results are similar to the other studies showing that Trump uses less diverse vocabulary and shorter sentences. One study looks for the linguistic changes in Trump's speech through time [56]. Their results support that he used more filler words as the time passed but his unique word count did not change. Political speeches are also analyzed to detect ideological styles, which are populist, elitist, or pluralist [57]. Another comprehensive study researches the traditional end of year addresses by nine presidents of the Italian Republic [58]. They use several statistical tests to identify distinctive words between presidents and within speeches of each president and report the results. Statistical tools are highly used in the studies and they are still the main tools for the analysis of spoken discourse. As a common point for these studies, it should be remembered that politicians usually deliver their speeches from a written text.

3.3 Dataset

Our dataset consists of interviews with 98 people from four big cities of Turkey, Ankara, Antalya, İstanbul, and Kocaeli. The interviews are conducted independently, in Turkish, and recorded via video camera. The participants are asked questions about being a moviegoer in the 60s, 70s, and 80s in Turkey. Later, conversations are transcripted by graduate students for each participant[1]. Figure 3.1 shows an example interview question and answer in Turkish and English. The reply shows the informal nature of the conversations. All the interview questions can be found in Figure 3.3 both Turkish and English. Two snapshots from an interview video are provided in Figure 3.2. A participant in his conversation, says that he calls a coffee mug "Türkân Şoray" due to color resemblance of the mug to Şoray's cloths in her movies with Cüneyt Arkın, and he shows a ticket from 1969 for Büyük Cinema in Ankara [1, p. 191]. The figure reflects that in the interviews nostalgia and ephemeral items are among the major themes.

[1]We plan to share the transcripts with and without interview questions on github.com/sevilcaliskan

İlk gittiğiniz filmi hatırlıyor musunuz?
Tabii ki hatırlıyorum, hatırlıyorum, hatırladığımı sanıyorum. İnci Sineması vardı. İnci Sineması belki hala vardır. Hukuk Fakültesi'nin yanında, hemen yanında. Hukuk Fakültesi'nin yanından bir sokak yukarı doğru çıkar, Cebeci'ye doğru. Ondan sonra hemen bitişiğinde olması lazım. Oraya gittim şeklinde hatırlıyorum. En küçük ablamla Sevim ablamla gittik. Ben herhalde ancak 3-4 yaşındaydım, ancak 3-4 yaşındaydım. Bir Neriman Köksal filmi diye hatırlıyorum ve ablamı şöyle yokladığımı hatırlıyorum "ablam karanlıkta gitti mi, gitmedi mi, gitti mi, gitmedi mi" şöyle herhalde on defa filan yokladıktan sonra artık yoklamamaya başladım, ondan sonra, Neriman Köksal iyi bir rolde değildi galiba böyle işte bir kötü kadın, o kötü kadın mı dedim, evet o kötü kadın dedi filan. Yani işte öyle İnci Sineması'nda bir Neriman Köksal filmi diye hatırlıyorum. Üç dört yaşındayken gördüm.

Do you remember the first movie you went to?
Of course I remember, I remember, I think I remember. There used to be İnci Cinema. Perhaps İnci Cinema still exists. Next to the Faculty of Law, right next to it. A street goes up, by the Faculty of Law, towards Cebeci. It must be right after that, next to it. I remember it like that I went there. We went with my youngest sister Sevim. I must have been only 3-4 years old, only 3-4 years old. I remember it being a Neriman Köksal movie and I remember that I was checking my sister "did my sister leave in the dark, did she stay, did she go?" I must have checked her for about ten times then stopped checking her, after that, Neriman Köksal was not in a good role, something like that, a bad woman, is she a bad woman I said, yes she's a bad woman she said, and so on. Well it's like that, I remember it as a Neriman Köksal movie in İnci Cinema. I saw it when I was three or four.

Figure 3.1: One example from interview questions.

(a) The mug named after Türkân Şoray.

(b) A movie ticket from 1969 for Büyük Cinema in Ankara.

Figure 3.2: Two snaphots from the interviews [1, p. 191].

1. Sizi tanıyabilir miyiz? İsminiz, hangi yılda doğdunuz, yaşınız?
   Can we get to know you? Your name, what year were you born, your age?

2. Anneniz, babanız ne iş yapardı?
   What did your mother, father do for a living?

3. Kardeşlerinizle ilişkileriniz iyi miydi, yakın mıydı yaşlar?
   Did you have good relationships with your siblings, were your ages close?

4. O dönemde toplumsal ilişkiler nasıldı, ne hatırlıyorsunuz çocukluğa dair?
   How were relationships in society at that period, what do you remember about childhood?

5. Ekonomik ve sınıfsal olarak nasıl tanımlarsınız o dönemleri?
   How would you describe those periods economically and in terms of classes?

6. Çocukluğunuzda politik ortam nasıldı?
   How was the political environment in your childhood?

7. Şu anda nerede yaşıyorsunuz ve ne işle meşgulsünüz?
   Where do you live now and what do you do for a living?

8. İlk gittiğiniz film neydi? Nerede, ne zaman ve kiminle gitmiştiniz?
   What was the first movie you went to? Where, when, and with who did you go?

9. Çocukken ne kadar sıklıkla sinemaya giderdiniz?
   How often did you go to the movies when you were a child?

10. Sinemaya kimlerle giderdiniz?
    Who did you go to cinema with?

11. Aileyle gittiğiniz filmlerle arkadaşlarınızla gittiğiniz filmler farklı mıydı? Türk filmleri ya da yabancı filmler ya da film türleri, o filmlerde farklılık olur muydu?
    Were the movies you went to with your family different from the movies you went to with your friends? Would there be differences in Turkish movies, foreign movies or genres?

12. Yazın yazlık sinema var mıydı?
    Were there open-air cinemas during the summer?

13. Sinemada ara verilir miydi, bir şey yenilip içilir miydi?
    Were there intermissions, would anything be eaten and drunk?

14. En keyif aldığınız sinema hangisiydi?
    What was the movie theater you enjoyed most?

15. Bilet fiyatları nasıldı 70-80'lerde? Pahalı diye mi hatırlarsınız yoksa normal mi diyorsunuz?
    How were ticket prices in the '70-'80s? Do you remember it as being expensive or normal?

16. Arkadaşlarla konuşur muydunuz, filmden çıktıktan sonra herhangi bir yere oturur muydunuz? Bir ritüeli var mıydı?
    Would you talk with friends, did you sit anywhere after leaving the movie? Was there a ritual?

17. Sinema dergilerini takip eder miydiniz?
    Did you follow the cinema magazines?

18. Televizyon evinize ne zaman geldi?
    When was the television introduced to your home?

19. Sinemaya hala gidiyor musunuz? Hangi türleri takip ediyorsunuz?
    Do you still go to the cinema? Which movie genres do you follow?

Figure 3.3: Questions asked during interviews in Turkish and English.

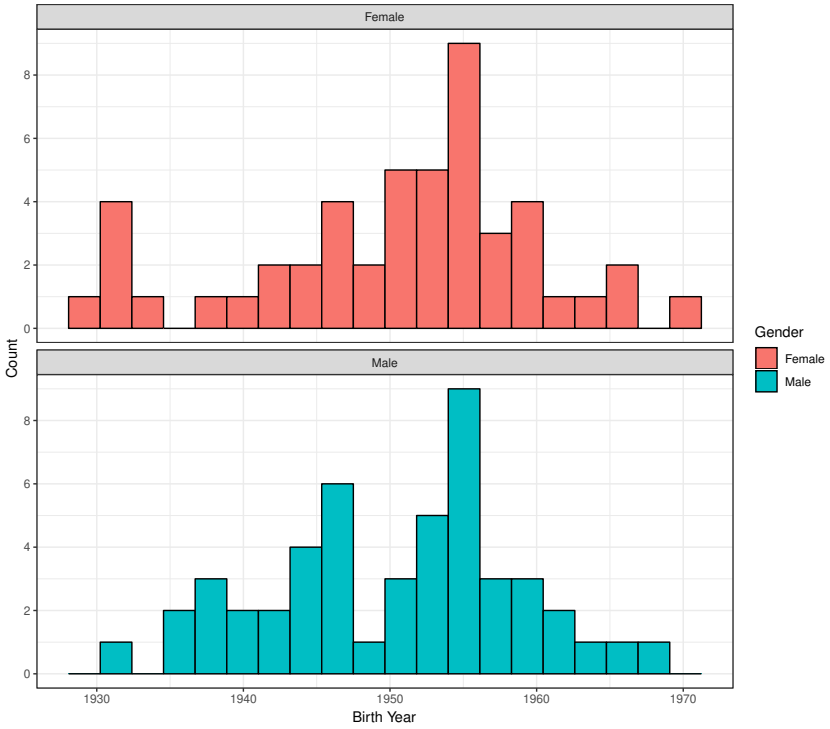


Figure 3.4: Histogram of the birth years of 98 participants.

There are an equal number of male and female participants, 49 of each. At the time of the interviews, their ages vary between 48 to 88, with a median age of 64, where the median age in Turkey in the year 2019 is 30.2 [59]. Figure 3.4 shows that the ages of each gender are similarly distributed. The occupations of participants are diverse, and include engineers, medical doctors, teachers, lawyers, academics, workers, tailors, housewives, etc. [1, p. 63]

The length of the interviews varies from 15 minutes to 330 minutes, with a mean 64.60 minutes and a standard deviation of 50.73 minutes, for total of 4,587 minutes (the lengths of some of the interviews are not known, so the given total is the total of known ones). It has a high variation and this variation is also reflected in word counts of each participant. The word count for each interview has a mean of 3,454.51 and a standard deviation of 1,834.07, and the total count is 336,018 words. Figure 3.5 shows the word counts of the speakers against their birth year. Figure shows that six participants have word counts less than 1,000, which caused their data to be eliminated from most of the experiments due to

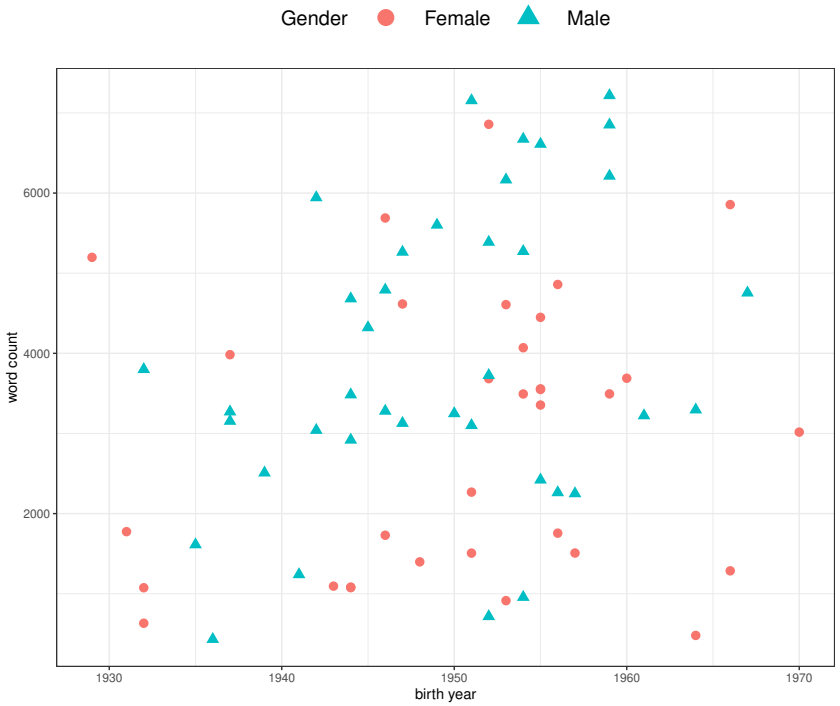


Figure 3.5: Word count versus birth year of participants.

shortness of their talk.

Figure 3.6 shows the duration against word count for each participant and they are correlated as expected. Histograms for each variable, duration, and word count, can be seen parallel to the axes of each variable. We see that interview duration mostly lays between 15 to 100 minutes, while the word count seems to have a more uniform distribution. This is probably due to the participants' different pace of talking.

Figure 3.7 shows the average word count per minute for each participant against the birth year of the participant. Smoothed lines show that older men speak slower than younger ones, however that is not the case for women. The speaking pace seems to be almost stable among different aged women. The duration also includes the time in which questions are asked, however it is of negligible importance since all the participants are asked almost the same set of short questions.

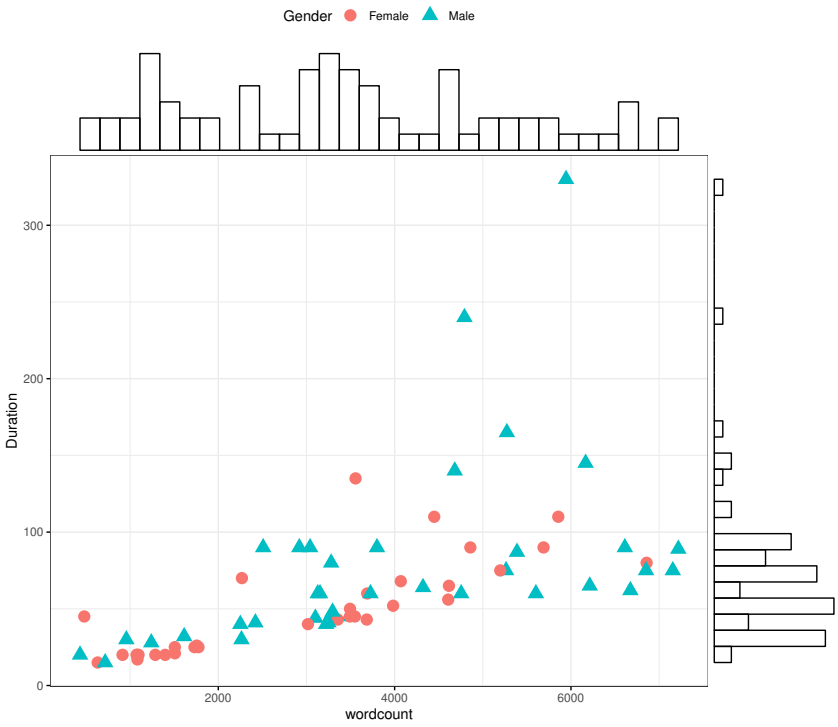


Figure 3.6: Duration versus word count of participants, bars indicate the frequency of observations.

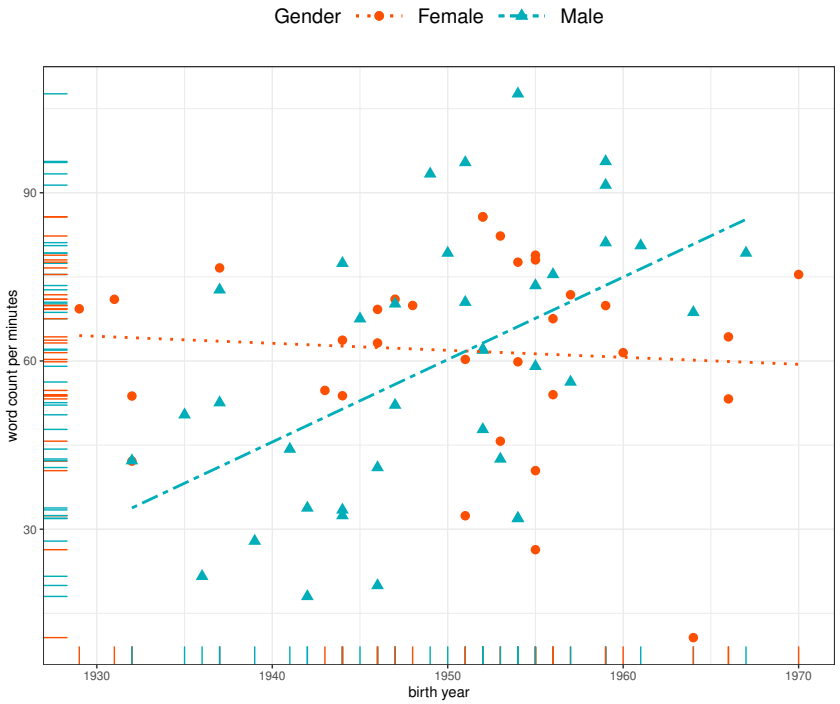


Figure 3.7: Average word count per minutes for each participant against birth year.

We want to see whether the change of duration of the interviews with age, that Figure 3.7 reflects, is significant. Linear regression is fitted without a gender variable and the result is insignificant with a low R-squared (-0.014) and a high p-value (0.8819). When we add gender as a variable to the model, however, we see that the p-value becomes somewhat significant (0.072) and the p-value for the gender variable becomes significant with p-value 0.023, indicating that gender affects the duration. The equation becomes:

$$Duration = 0.026 * birthYear + 27.56 * genderMale - 1.61 \tag{3.1}$$

In this formula, the $genderMale$ variable is either 1 or 0, depending on the gender of the participant. The coefficient of the variable gender male is 27.56, which means male participants talk about 27 minutes longer than female participants on average, which is contrary to the belief that women talk more.

3.4 Research Questions and Experimental Design

3.4.1 Research questions

The experimental environment is set upon the following research questions:

RQ1: *Can the age and gender of the participant be inferred from their speech as is the case for written text?*

RQ2: *Are there any differences in language use between younger and older people?*

RQ3: *Can we identify a text as being either spoken or written?*

RQ4: *Can we predict the authorship of spoken text given the person's written text?*

Other research questions are evoked by the data. One can further investigate whether other kinds of demographic information regarding the participant can be inferred from the speech. This information may be about the education level of the participant, occupation, and the nature of the city that the participant lives in such as a big city or a small town. For this study, all of the participants do not possess all demographic attributes as it is offensive to directly ask for that information. In some conversations, that information is given by the participant voluntarily, but the number is not sufficient to conduct experiments.

Participant identification is another problem that can be studied with the dataset. However, in this study, we are interested in the issues related to participant (author) profiling [9] and language use. The issues related to participant identification and verification are left as problems we want to study in our future research.

3.4.2 Experimental design

The experiments include the analysis of participants through their speeches. Interviews are analyzed stylometrically, by using different sets of stylometric features for different groups of participants. The gender group includes two genders, male and female. Age groups are formed with participants whose birth years fall into the same decade, out of the four decades through 1930 and 1960.

Figure 3.8 shows the 30 most frequently used words among all the participants; Table 3.1 shows the first 10 most frequently used words for each age and gender group. Table 3.1 shows that the frequency of used words changes among groups. With the help of these variations, we aim to detect the age and gender group of a participant. To overcome the differences in the lengths of speeches and to strengthen the power of our analysis, we generate word blocks of similar lengths from interviews. For the sake of not splitting sentences, we generate word blocks having the given number of words within a complete sentence. As a result, the lengths of word blocks are close but not identical. In order to generate an equal number of blocks from each group, we pick a random sentence from a speech as

bir (a, an, one), o (he, she, it), çok (very), yani (so), da (too), de (too), vardı (there was, there were), ama (but), daha (more), ve (and), ben (I), sonra (later), şey (thing), bu (this), zaman (time), film (movie), için (for), işte (here is), sinema (cinema), mesela (for example), böyle (so), var (there is, there are), öyle (so), gibi (like), şimdi (now), falan (like), biz (we), sinemaya (to the cinema), ki (that, which, who), ne (what).
ama (but), ben (I), bir (a, an, one), biz (we), böyle (so), bu (this), çok (very), da (too), daha (more), de (too), falan (like), film (movie), gibi (like), için (for), işte (here is), ki (that, which, who), mesela (for example), ne (what), o (he, she, it), öyle (so), sinema (cinema), sinemaya (to the cinema), sonra (later), şey (thing), şimdi (now), var (there is, there are), vardı (there was, there were), ve (and), yani (so), zaman (time).

Figure 3.8: Top 30 most frequently used words by all participants ordered by decreasing order of frequency and alphabetically.

Birth Decade	Gender	Most Frequently Used 10 Words (ordered)
1930s	F	bir, o, çok, yani, da, de, ama, vardı, ben, öyle
	M	bir, çok, ve, o, vardı, yani, ama, sonra, da, iyi
1940s	F	çok, bir, o, da, de, yani, sonra, ben, vardı, ama
	M	bir, o, yani, da, çok, de, vardı, ama, sonra, bu
1950s	F	bir, çok, o, yani, da, vardı, de, ama, daha, ben
	M	bir, o, çok, yani, da, de, vardı, ve, ama, sonra
1960s	F	bir, o, çok, yani, da, de, ama, daha, vardı, sonra
	M	bir, o, yani, çok, da, de, bu, şey, ama, ben

Table 3.1: Most frequently used words of different age and gender groups of participants.

Statistical style markers	Example											
Frequencies of most frequently used words	bir 225	o 96	çok 109	yani 26	da 56	de 50	ama 51	vardı 36	ben 63	ve 1110		
Statistics	Sentence 242	Type 1127	Token 1995	Char 11322	Type / token 0.56	Hapax L. 824	Dis L. 166					
Word length frequencies	1 10	2 54	3 115	4 225	5 400	6 398	7 452	8 397	9 316	10 219	11 113	12 79
Suffix frequencies	A1sg 66	A2sg 12	A3sg 1087	Fut 2	PastPart 22	Pres 35	PresPart 18	Able 1				

Table 3.2: Examples of stylometric features of a word block, length of 2,000 words.

the starting sentence of a block and add sequential sentences until the required word length is met. This way, we can pick random parts from texts. This set up lets us generate word blocks of equal number for each group and augment data for the groups with a fewer number of sentences.

After generating blocks, we obtain stylometric features for each block and use the features for experiments. We choose the most used style markers in the literature [9], and keep as many as possible, following Rudman's suggestion that different style markers should be studied [60]. We also include suffix frequencies to see if different age and gender groups use different tenses and person suffixes [61]. We conducted experiments with several combinations of different style markers. Those experiments produced similar accuracy results when certain style markers are included. As a result, we find grouping them to be most useful.

The style markers are grouped into four categories: 1. frequencies of most frequently used words, 2. some statistics of a word block, 3. word length frequencies, and 4. suffix frequencies. Table 3.2 shows the examples of the feature groups for a word block whose length is 2,000 words. The most frequently used words list is generated by counting word appearances in each group/class, picking the top N most frequent words of each, and unifying them into a word list. We also generate

Short: Long	Short: Long
A1pl: First Person Plural	JustLike: Just Like
A1sg: First Person Singular	Loc: Locative
A2pl: Second Person Plural	Ly: Ly
A2sg: Second Person Singular	Narr: Narrative Tense
A3pl: Third Person Plural	NarrPart: Narrative Participle
A3sg: Third Person Singular	Neces: Necessity
Abl: Ablative	Neg: Negative
Able: Ability	Ness: Ness
Acc: Accusative	Opt: Optative
Acquire: Acquire	P1pl: First Person Plural Possessive
AfterDoing: After Doing	P1sg: First Person Singular Possessive
Agt: Agentive	P2pl: Second Person Plural Possessive
Aor: Aorist	P2sg: Second Person Singular Possessive
AorPart: Aorist Participle	P3pl: Third Person Plural Possessive
AsIf: AsIf	P3sg: Third Person Singular Possessive
AsLongAs: As Long As	Pass: Passive
Become: Become	Past: Past Tense
ByDoingSo: By Doing So	PastPart: Past Participle
Caus: Causative	Pres: Present Tense
Cond: Condition	PresPart: Present Participle
Cop: Copula	Prog1: Progressive1
Dat: Dative	Prog2: Progressive2
Desr: Desire	Recip: Reciprocal
Dim: Diminutive	Reflex: Reflexive
Equ: Equ	Rel: Relation
Fut: Future	Related: Related
FutPart: Future Participle	SinceDoingSo: Since Doing So
Gen: Genitive	When: When
Imp: Imparative	While: While
Inf1: Infinitive1	With: With
Inf2: Infinitive2	Without: Without
Inf3: Infinitive3	Zero: Zero
Ins: Instrumental	WithoutHavingDoneSo: Without Having Done So

Table 3.3: Abbreviation list of suffixes that are counted in experiments [2].

the same features with the most frequent bi-grams, which are sequences of two adjacent words; however, experiments with the bi-grams were not as successful as the experiments with other feature groups, so we do not share them. In Table 3.2, an example of the vector with the first 10 most frequent words can be seen in the first row.

The statistics row of Table 3.2 shows the calculated style markers for each word block as another group. The features in this group are sentence count, type count, which is the number of distinct words, token count, which is the number of all words, character count, type-to-token ratio, which is an indicator of vocabulary richness, hapax legomenon, which is the number of words appearing only once, and dis-legomenon, which is the number of words used twice, type and token character count, average token and type lengths and variations, and average sentence length and variation. The word length frequencies row represents the frequencies of the words with given word lengths, and goes up to the longest word length in the training set. Finally, suffix frequencies represent the frequencies of different suffixes in the speech, which are extracted by the Zemberek API [2]. *A1sg* is the tag for first-person singularity suffix, which is added to the verbs in Turkish. *Fut* means future tense. *Able* is for the ability suffix. Detailed information regarding the suffixes and abbreviations used can be reached from [62]. Please note that we are not able to present all of the features in Table 3.2, and provided ones are only a sample, for the purpose of illustration. The list of all the suffixes can be found in Appendix 3.3.

As a classifier, we use Support Vector Machines (SVM) [63], since, in our experiments, it gives the most successful results among logistic regression, discriminant analysis, and neural networks with word embeddings (both recurrent and multilayered). SVM is applied with 5 times Monte Carlo cross-validation for each individual experiment, meaning we split train and test data randomly for each of 5 experiments and average their results. We use 80% of the data as the training set and 20% as the test set.

Word block length	Number of generated blocks
20	9800
50	2940
100	1470
500	475
1000	172
2000	71

Table 3.4: Total blocks generated for each block length parameter for gender classification experiments (the same number of blocks generated for each person).

3.5 Experimental Results

In this section of the chapter, we provide the experimental results for our research questions in the order they were asked.

3.5.1 RQ1: Predicting age and gender of the participant

In this section, the task is to distinguish the age and gender of the participant from their speech. A similar task gives successful results when applied to the written texts. For example, out of 40 Turkish novels, the gender of writers is classified correctly with an accuracy of 88.23%, while female authors classified with 100% accuracy [7]. In the same study, the era of the novel, in which it is written,rather than the age of the authors, is also predicted with 57.27% accuracy, out of 4 eras, where each is composed of 25 years. We anticipate to observe similar patterns in spoken language as well.

For gender classification, we generate the same number of blocks for each person. Table 3.4 shows the total number of blocks generated for different word lengths. As the length of the blocks increases, we decrease the number of blocks we use to avoid generating similar blocks. If the length of a speech is shorter than the intended block length, we do not generate blocks from that speech. Therefore, changes in the number of blocks are not consistent.

	Number of most frequently used words					
Word block length	10	25	50	100	150	200
20	0.60	0.67	0.73	0.77	0.80	0.80
50	0.61	0.69	0.73	0.77	0.77	0.79
100	0.64	0.70	0.74	0.81	0.82	0.82
500	0.73	0.84	0.86	0.91	0.92	0.90
1000	0.75	0.83	0.82	0.84	0.87	0.88
2000	0.68	0.74	0.65	0.81	0.79	0.65

Table 3.5: Accuracy results of classifying gender with different number of most frequently used words and word blocks.

Figure 3.9: PCA plot of the gender of the participants using top 150 most frequent words and 5 blocks for each participant.

	Word block length					
Features	20	50	100	500	1000	2000
Statistics	0.56	0.59	0.62	0.61	0.62	0.65
Word length frequency	0.58	0.60	0.63	0.63	0.64	0.68
Suffixes	0.61	0.64	0.64	0.76	0.74	0.70

Table 3.6: Accuracy results of classifying gender with different word block lengths and feature groups.

The results of the experiments with the generated word blocks are given in Table 3.5 and Table 3.6. Table 3.5 shows the accuracy results of the gender classification experiments completed with different numbers of most frequent words and different lengths of word blocks using SVM. We see that as the number of most frequent words increases, accuracy usually increases, since in such a case, the vector approximates to the bag-of-words model. As the word block length increases, accuracy, also, generally increases since it is easier to catch patterns in longer texts. We see that when the word block length is 2,000, accuracy values decrease since the number of generated blocks is decreased to a point where it affects prediction accuracy. The highest accuracy values are reached when the word block length is 500 and the numbers of most frequently used words are 100 or 150. The same experiments are performed by assigning gender labels randomly, and we see that the results of the experiments are significantly better than the random case, which is 50% or lower for almost all cases. Table 3.6 shows the experimental results with other feature groups. Accuracy is are highest when the experiments conducted with the most frequently used words.

A principal component analysis (PCA) plot is generated with the parameters of the experiment with the highest accuracy, which is the experiment with the first 150 most frequently used words and blocks of 500 words, with 5 blocks for each participant. Each point in Figure 3.9 represents the word block vectors from the experiment, and the two dimensions in the plot show only 6.2% of the variation of the features. Even with such a low percentage, we can observe the separation between male and female blocks.

Word block length	Number of generated blocks
20	4000
50	2000
100	800
500	400
1000	200
2000	80

Table 3.7: Total blocks generated for each block length parameter for age group classification experiments (the same number of blocks / samples generated for each decade class).

	Number of most frequently used words					
Word block length	10	25	50	100	150	200
20	0.36	0.47	0.45	0.55	0.60	0.60
50	0.51	0.58	0.63	0.68	0.69	0.70
100	0.47	0.57	0.62	0.65	0.67	0.66
500	0.64	0.75	0.78	0.84	0.84	0.85
1000	0.67	0.75	0.85	0.81	0.84	0.87
2000	0.69	0.69	0.71	0.75	0.69	0.78

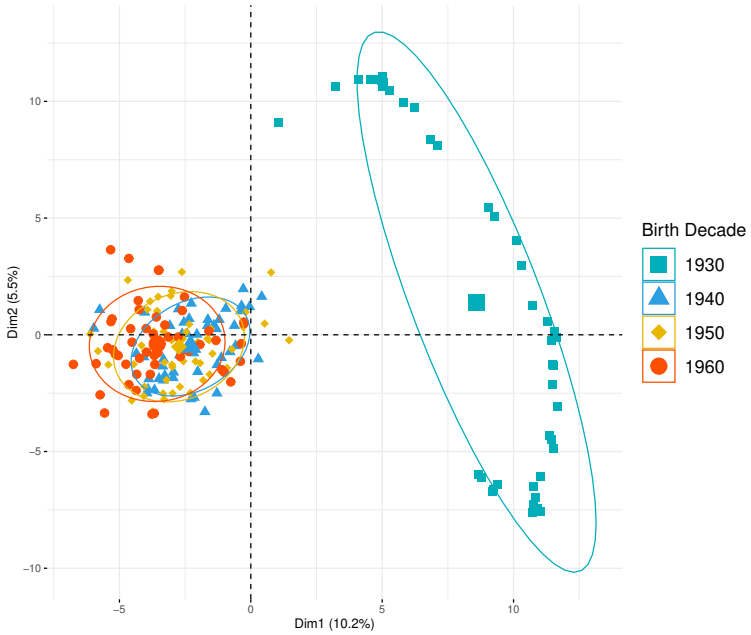
Table 3.8: Accuracy results of classifying birth decade with different number of most frequently used words and word blocks.

	Word block length					
Features	20	50	100	500	1000	2000
Statistics	0.35	0.40	0.40	0.60	0.61	0.58
Word length frequency	0.37	0.46	0.47	0.60	0.60	0.65
Suffixes	0.34	0.38	0.48	0.64	0.75	0.64

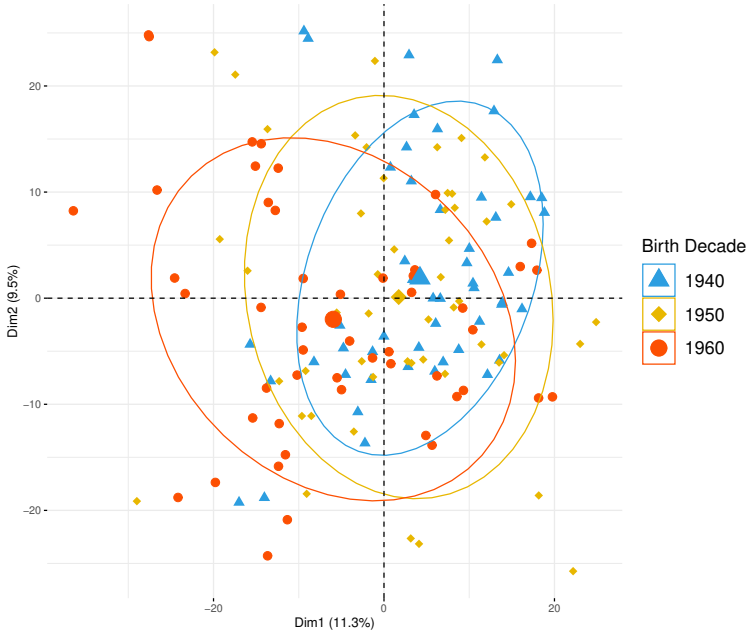
Table 3.9: Accuracy results of classifying birth decade with different number of word block lengths and different feature groups.

Age group	Number of words
1930	19,050
1940	75,068
1950	153,521
1960	88,379

Table 3.10: Number of total words in each age group of subjects.



(a) PCA plot of the speeches with all age groups.



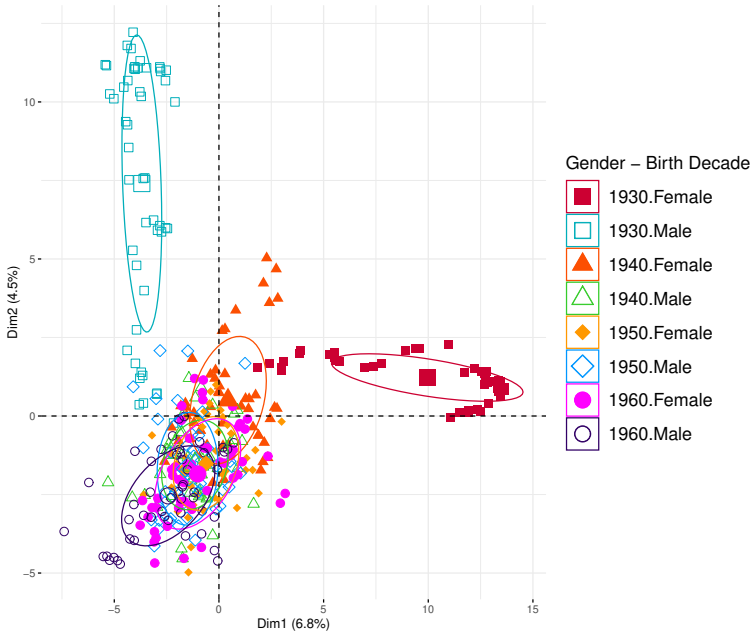
(b) PCA plot of the speeches after the 1930 group is removed.

Figure 3.10: PCA plot of the speeches showing birth decade groups with the top 200 most frequent words and 50 blocks of length 1,000 for each group.

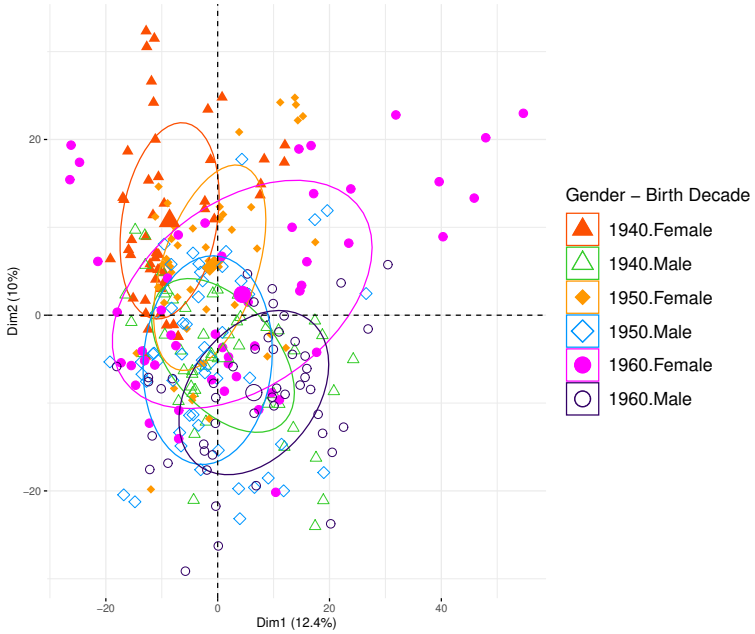
For age classification, we generate the same number of blocks from the concatenated text of each age group, since some of the age groups have a smaller number of participants. Table 3.7 shows the total number of blocks generated for different word lengths. The number of generated word blocks is smaller since some age groups have much shorter texts than others, as given in 3.10, and the number of generated blocks are picked to avoid overlapping text blocks. The results of the experiments with the generated word blocks are given in Table 3.8 and Table 3.9. Table 3.8 shows the accuracy results of the age group classification experiments completed with different numbers of most frequent words and different lengths of word blocks. Similar to gender classification experiments, as the number of most frequently used words increases, accuracy usually increases. As the word block length increases, accuracy also increases. The highest accuracy values are reached when word block length is 1,000. Again, the same experiments are applied by assigning age group labels randomly; we see that random success is around 25% which is much less than the success of the experiments. Table 3.9 shows the experimental results with other feature groups. We see that accuracy results are not as high as the experiments conducted with most frequently used words, similar to gender experiments.

Figure 3.10 shows the PCA plot generated with blocks of length 1,000 and the top 200 most frequent words the experiments. Each point represents the word block vectors. We see that the 1930 group is different from the other groups, and stands out. One reason for the separation may be that the number of participants from both genders who fall into the 1930s group is less than other age groups. Since we generate the same number of word blocks for each age group, we might say that data of the 1930s group can be biased towards the subjects' data. To observe the difference between other groups, we remove the 1930 group and plot the blocks again. Some separation of the groups can be seen with two dimensions reflecting around 20% of the variation and it is noticeable that age groups are clustered from left to right in the order of 1960, 1950, 1940, and 1930.

Another set of experiments is conducted by grouping both age and gender together, in which we end up with 8 different groups. The number of generated blocks is the same as the age groups as in the Table 3.8, again blocks are generated



(a) PCA plot of the speeches with all groups.



(b) PCA plot of the speeches after the 1930 group is removed.

Figure 3.11: PCA plot of the speeches showing age and gender groups with the top 150 most frequent words and 50 blocks of length 1,000 for each group.

	Number of most frequently used words					
Word block length	10	25	50	100	150	200
20	0.31	0.37	0.51	0.57	0.61	0.62
50	0.40	0.53	0.60	0.66	0.71	0.53
100	0.36	0.50	0.59	0.67	0.66	0.68
500	0.63	0.81	0.86	0.91	0.89	0.88
1000	0.69	0.85	0.86	0.90	0.89	0.86
2000	0.72	0.78	0.75	0.81	0.82	0.57

Table 3.11: Accuracy results of classifying birth decade and gender together with different numbers of most frequently used words and word blocks.

	Word block length					
Features	20	50	100	500	1000	2000
Statistics	0.20	0.24	0.24	0.39	0.48	0.52
Word length frequency	0.22	0.31	0.26	0.51	0.53	0.61
Suffixes	0.25	0.33	0.37	0.62	0.78	0.83

Table 3.12: Accuracy results of classifying birth decade and gender with different numbers of word blocks lengths and different feature groups.

by concatenating speeches from the same groups. Results of the experiments are given in Table 3.11 and Table 3.12. Table 3.11 shows the accuracy results of the age group classification experiments completed with different numbers of most frequent words and different lengths of word blocks. Given that the random success for this experiment group is around 12.5%, results are better than gender (Table 3.5 and 3.6), and age (Table 3.8 and 3.9) groups classifications. Table 3.12 shows the experimental results with other feature groups and again, accuracy results are not as high as the experiments conducted with the most frequent words. These results suggest that there may be an interaction effect between age and gender. Different aged men and women have different key patterns or words that they use, hinting their age and gender group.

Figure 3.11 shows the PCA plot, in which each point represents the word block vectors. We again observe that the 1930 group is different from the other groups, and it stands out for both genders. To see the difference between other groups, we remove the 1930s group and replot the blocks. Some separation of groups can

be observed within two dimensions reflecting, around 25% of the variation of the features. The separation between male and female groups can also be observed.

3.5.2 RQ2: Differences in language use between the young and old

In this section of the chapter, the task is to determine the differences in language use among age groups. The experiments of RQ1 show that language use between young and old people differs to some extent. We conduct further experiments on the dataset by dividing the subjects into four age groups.

First, we want to see the differences in the most frequent word usage in age groups. To see if the differences are significant, we choose the set-up of the most successful experiment at separating age groups. We generate 20 blocks for each group, which are 1,000 words in length. 20 is chosen as the parameter for the number of word blocks to avoid overlapping in the blocks since the first age group has fewer words than others. The number of words in each group can be found in Table 3.10. We use the first 20 most frequent words for the analysis for simplification purposes. We use the permutational multivariate analysis of variance (PERMANOVA) for analysis [64]. It is chosen as the analysis tool since the analysis is multivariate; the 20 most frequent words and response variables, which are the counts of words in text blocks, do not meet the assumptions of multivariate analysis of variance (MANOVA). We use Euclidean distance for PERMANOVA analysis and use the Holm method [65] for the correction of pairwise comparisons. PERMANOVA gives a p-value of 0.001, which indicates that the most frequent words usage differs significantly between age groups. To see the pairwise differences, we apply pairwise tests with correction. Table 3.13 shows the p-values of pairwise analysis; we see the age groups also differ significantly in pairwise comparisons, except for the 1940s and 1950s groups, which seem to be the most similar.

Table 3.14 shows the words whose usage mostly differs between age groups. We

	1930	1940	1950
1940	0.006	-	-
1950	0.006	0.014	-
1960	0.006	0.006	0.006

Table 3.13: p-values of pairwise PERMANOVA analysis.

Age groups	Most frequent words
1930 - 1940	çok, bir, yani, da, öyle, o, ve, ben, film, sonra, sinema
1930 - 1950	çok, bir, da, öyle, yani, vardı, o, ve, daha, ama, tabi
1930 - 1960	bir, çok, yani, da, ve, o, vardı, öyle, zaman, daha, de
1940 - 1950	yani, bir, çok, vardı, da, o, ve, daha, film, ben, sonra
1940 - 1960	yani, bir, çok, o, vardı, da, ve, daha, ben, öyle, de, film
1950 - 1960	yani, bir, vardı, çok, o, da, ve, de, öyle, tabi, ben, daha

Table 3.14: Most frequent words whose usage varied the most among age groups.

reconduct the same experiments by also adding gender as a dependent variable. The differences are again significant in a gender-aware set-up and all p-values are less than 0.05. For brevity, they are not given.

Earlier studies compared written language change in Turkish with time [7, 8]. They used linear regression to see the differences in token and type lengths among new and old Turkish literature texts. They show that average token and type lengths increase as the age of works decreases, meaning in Turkish written language, words get longer as time passes. To see if this is similar in spoken language in terms of the age of the participant, we test the hypothesis that average token and type lengths increase with the increase of the birth year of the participants. We use linear regression as well, by generating only one text block for each participant, with 1,000 words length. The response variable becomes the average token or type length of the subject, and the independent variable is the age. For both linear regression analyses, we see that the p-value of the coefficient of the age is much higher than 0.05, and the p-value of the constant is less than 0.01, so we reject the hypothesis that token and type lengths change with the age.

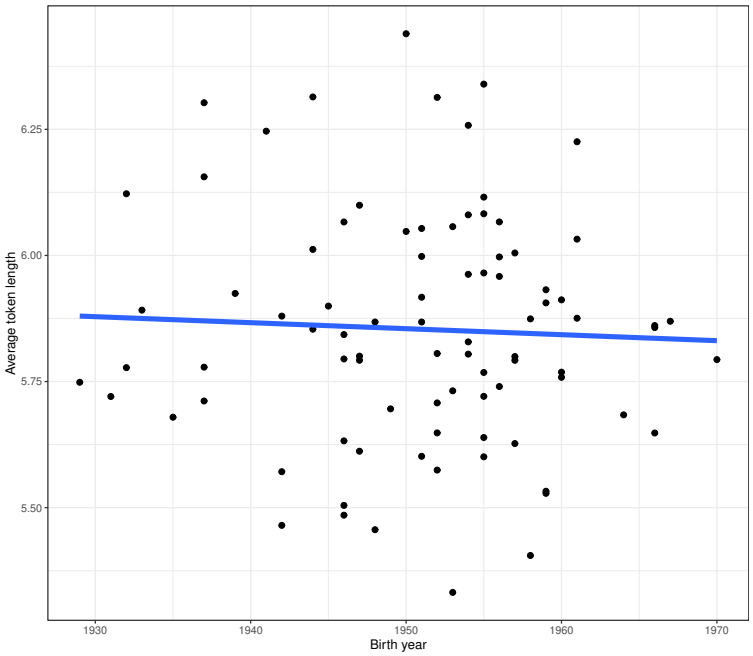
We observe an increase in type length as the birth year increases, hinting an observation similar to the 20th century Turkish literature results; however, from the regression results we understand that this change is statistically insignificant. Figure 3.12 shows the scatter plot for both average token and type lengths against age. The results are probably due to participants adjusting their spoken language according to the current vocabulary and language style while written texts reflect the literary language of their time.

We also analyze age groups in terms of vocabulary richness using type to token ratio, hapax legomenon, and dis legomenon with linear regression analysis. We, again, see that p-values for the coefficients of all the metrics are higher than 0.05, and the p-value of the constant variable is much less than 0.01, showing that vocabulary richness does not differ significantly between age groups.

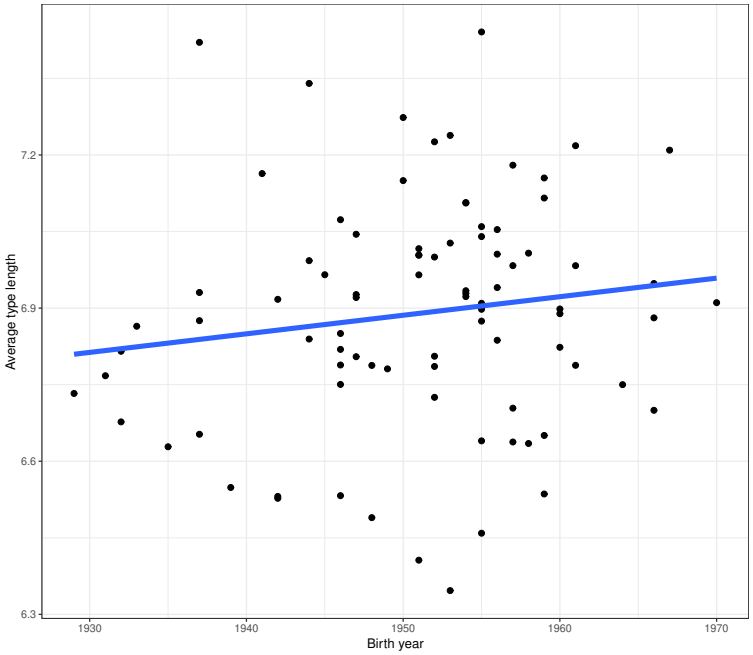
3.5.3 RQ3: Identifying a text as spoken or written

In this section of the chapter, the task is to categorize a given passage as spoken or written using the mentioned subjects' both interviews and written texts. Note that some of the subjects of the interviews are academics and writers, and have some written pieces of their own. Biber [42] made a similar analysis using factor analysis, focusing mainly on the differences in parts of speech. In another study, written and spoken language samples of the subjects are analyzed and results show that written texts are usually shorter than spoken texts; they have longer words and a higher variety of vocabulary [66]. In this section, we use the most frequent words, frequency statistics of the texts and distributions of word length frequencies, and suffix frequencies as same as the previous sections.

We use the written and spoken texts of 5 subjects, whose written texts are available. The written corpus is the text that we were able to find for these subjects. Table 3.15 shows the number of words in both written and spoken texts for each subject. We see that for subject 3, the number of words in written text is less than 1,000. So, we choose the largest word block length as 700, so as to be able to generate at least a couple of blocks of the subject. Table 3.16 shows the



(a) Scatter plot of average token lengths.



(b) Scatter plot of average type lengths.

Figure 3.12: Scatter plots of type and token lengths, in terms of number of letters, against age of the subjects.

Subject	Number of words Written	Spoken
1	1,073	5,349
2	8,236	7,152
3	936	4,322
4	1,499	1,613
5	6,592	4,613

Table 3.15: Number of words each subject has in written texts and transcripts of interviews.

Word block length	Number of generated blocks
20	500
50	300
100	200
500	100
700	50

Table 3.16: Total blocks generated for each block length parameter for written-spoken text classification experiments (the same number of blocks / samples generated for each person).

	Number of most frequently used words					
Word block length	10	25	50	100	150	200
20	0.89	0.90	0.90	0.91	0.95	0.94
50	0.92	0.95	0.88	0.92	0.92	0.93
100	0.95	0.97	0.97	0.96	0.96	0.94
500	1.00	1.00	1.00	0.99	0.99	0.96
700	1.00	0.98	1.00	0.96	0.98	0.96

Table 3.17: Accuracy results of classifying texts as spoken or written.

	Word block length				
Features	20	50	100	500	700
Statistics	0.80	0.87	0.86	0.99	1.00
Word length frequency	0.81	0.84	0.88	0.98	1.00
Suffixes	0.78	0.82	0.86	1.00	0.96

Table 3.18: Accuracy results of classifying text as spoken or written with different numbers of word blocks lengths and different feature groups.

total number of generated blocks for each block length for different experiments. We, again, use SVM for classification purposes with the given setup parameters in Section 3.4.2.

Table 3.17 shows the results of the experiments with the most frequent words as features. We see that in many of the experiments, written and spoken language can be differentiated with high accuracy and some even with 100%. Table 3.18 shows the results of the experiment conducted with other feature groups. We see that as the length of the word block increases, spoken and written texts are classified with higher accuracy. PCA plot in Figure 3.13 is created with the most frequent words; the distinction between spoken and written word blocks is easily seen. The results indicate that written and spoken languages are indeed different, even if the interview participants and writers are the same. This can be explained by the fact that the nature of the speeches is casual and spontaneous while the authors of the written texts are generally academics, whose texts are of a more formal nature.

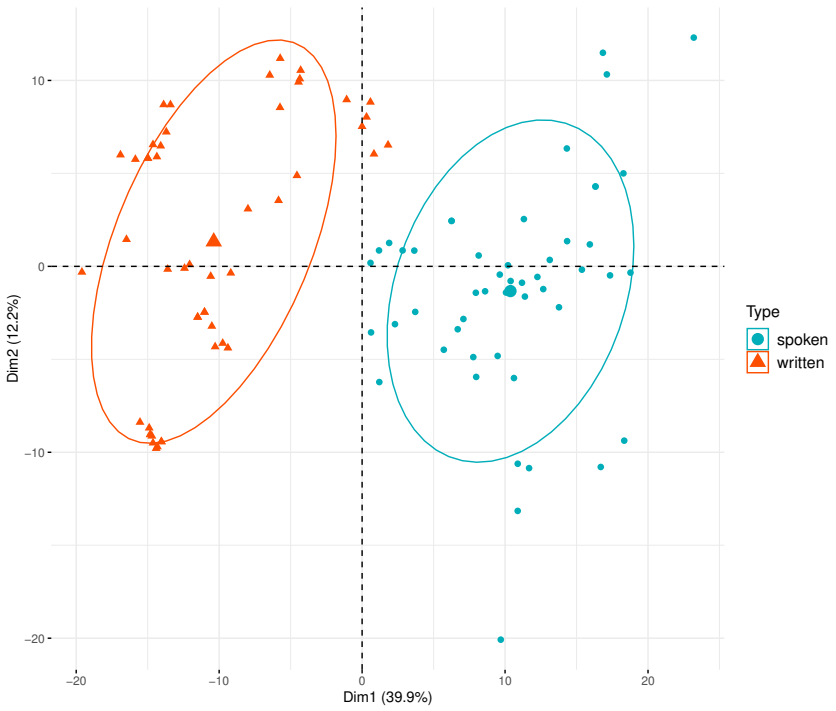


Figure 3.13: PCA plot of the spoken and written text blocks using top 20 most frequent words and 500 words length 10 blocks for each person.

3.5.4 RQ4: Predicting authorship of spoken text given the subjects' written text

In this section of the chapter, the task is to predict the author of a written text by looking at the author's spoken text, or vice versa. This task is also considered to be cross-genre authorship attribution and it is a challenging task since texts come from different genres and thematic areas [67]. In our case, as the spoken texts for all the subjects are on the same topic, it is even more challenging. Since, we know that most subjects are middle or upper-middle-class [68], their lifestyles, habits, and, probably, speaking styles and word choices in daily life are similar.

In the first set of experiments, we use our experimental setup with SVM and word blocks using different features. We generate the same number of text blocks as in Section 3.5.3, given in Table 3.16. The only difference is the label of the data, now being authors rather than the type of the text. We use written texts as training data and spoken texts as test data. In this setup, since we generate

	Number of most frequently used words				
Word block length	2	8	15	20	29
20	0.21	0.20	0.18	0.24	0.25
50	0.20	0.17	0.21	0.21	0.23
100	0.23	0.19	0.16	0.24	0.25
500	0.26	0.14	0.18	0.17	0.27
700	0.29	0.18	0.20	0.23	0.26

Table 3.19: Accuracy results of classifying the author of spoken texts from written texts with SVM.

	Word block length				
Features	20	50	100	500	700
Statistics	0.21	0.20	0.25	0.25	0.24
Word length frequency	0.24	0.25	0.23	0.18	0.22
Suffixes	0.17	0.22	0.17	0.16	0.16

Table 3.20: Accuracy of differentiating the author of spoken texts from written texts with different feature groups.

the same number of blocks for each text type of each author, train and test data amount becomes equal meaning 50% is used for training, and 50% is for testing. Table 3.19 shows the experimental results using SVM with the most frequent words and different parameters. We see that for some of the experiments in Table 3.19, accuracy is not better than random case of 20% and for some only slightly better. To see if the results better than random prediction are due to chance, we calculate the 95% confidence intervals for the accuracy of each experiment with different parameters. Many of the 95% confidence intervals, however, include the 20% success rate, meaning that slightly better results might be due to chance. Similarly, Table 3.20 shows the accuracy results for the experiments with the other features. We, again, see some slightly better-than-random results and some worse ones. 95% confidence intervals show that the better results are insignificant. Additional experiments are conducted by training SVM with spoken texts and testing with written texts, to predict the author of the written texts with the spoken texts. The results are similar to the ones above and not significantly better than the random guess, so we do not share them.

	Subject ID (spoken text)				
Subject ID	1	2	3	4	5
1	0.79	0.83	0.83	0.87	0.80
2	0.74	0.72	0.74	0.79	0.71
3	0.85	0.85	0.84	0.88	0.85
4	0.74	0.75	0.73	0.77	0.75
5	0.74	0.76	0.77	0.81	0.73

Table 3.21: Distances of subjects' written texts to spoken texts (lowest values are underlined).

Distance between texts is another method used for authorship attribution purposes [69]. It is used by attributing the author whose text is close or similar to the target text. We employ the Labbé's distance measure that is designed to get a normalized metric regardless of the length of two compared texts [70]. This way, we do not need to generate word blocks to get the distances between texts, also texts can be considered as a whole without any information loss. The Labbé distance measure is between 0.0 and 1.0, 0.0 meaning the compared texts are the same, and 1.0 meaning completely different texts. According to this measure, it is suggested that to attribute a text to an author, distances between the target text and the author's text should be less than 0.65. Table 3.21, shows the calculated distance measures of written texts of subjects to all other spoken texts. We observe that all of the measures are greater than 0.65, meaning we cannot attribute any written text to any participant. However, we also observe that for Subject 1, Subject 3 and Subject 5, their written text is the one closest to their spoken text. Although the distances are very close to the other spoken texts, it can still be said that, there are some weak similarities between written and spoken texts of a subject.

3.6 Conclusion and Future Work

In this chapter, we provide a comprehensive quantitative analysis of spoken discourse for the Turkish language using memoirs of a group of old-time moviegoers.

The aim is to infer various attributes of the participants from their speech. Our experiments show that, to a large extent, age and gender groups can be attributed. We can accurately classify a text as written or spoken. Also, we see that the spoken language of a person changes itself to fit into the current spoken language of society, regardless of the age of the participant. Language use among age or gender groups differs in terms of the frequencies of the used words, rather than the words. Attributing participants to spoken texts from written texts is a challenging task. The classification accuracy in the experiments is not higher than a random guess. This shows that more information or mining is needed for the task.

There are some future research possibilities that can be studied following this study. Additional language features can be investigated for analyzing groups. For example, n-gram usage differences can be further analyzed among age and gender groups. Other methods can be used for classification. Ensemble methods are more likely to give higher accuracy results, since they provide a classifier stronger than their components [71]. Movie choices of certain social, age, and gender groups can be investigated. Moreover, a social network can be created with the help of movie and movie theater names and the social networks of different groups can be compared. This may help us make sense of group choices and group dynamics of the past. Investigating the cultural heritage aspects of the memories is another path of research waiting to be done.

Chapter 4

Zero-shot Cross-Lingual Transfer Assessment from English to Turkish for Classification

4.1 Introduction

Given the significant improvements in NLP domain in the last years, researchers now use multi-lingual data together. Many studies are invested in the topics of multi-lingual and zero-shot learning with promising results [14–16]. Since many language suffer from the lack of data for specific tasks, such studies encourage researchers to use methods that can benefit data of multiple languages. multi-lingual data is used for many tasks like natural language inference, machine translation, classification and etc [17]. In this study, we focus on the problem of micro-text classification with limited or no data in a given language.

Using data from another language for training purposes, is a common and applicable solution, where data collection is not possible. Most proposed method is to translate the data of languages to a mediator language or one to another [18, 19]. However, it may cause information loss in data and is not practical since

translating adds an extra step to classification. Another method is mapping word embedding of to languages to the same space, which allow transferring from one language to another [20–22]. With the lately proposed, large and multi-lingual networks, zero-shot cross-lingual cross-lingual transfer can be used for similar purposes [14], Zero-shot transfer simply means training a model in a source language, and transferring to a target language. Models like BERT and XLM-R are pre-trained with data of many languages together and allow zero-shot learning with impressive performances on multi-lingual cases [23, 24].

In this chapter, we use multiple Turkish micro-blog datasets for classification. We use English micro-blog datasets from the same and different domains, to investigate zero-shot transfer performance of multi-lingual networks, BERT and XLM-R, on Turkish data. Our contribution with this chapter is, to assess the zero-shot cross-lingual abilities of multilingual networks and compare performances of the multilingual networks with newly released monolingual networks for Turkish in the classification tasks.

4.2 Related Work

There are many methods proposed to map word embeddings of different languages to the same space [72–74]. Many methods propose learning mappings independently with monolingual corpora, then aligning the embedding spaces with parallel data. Such methods are not suitable for low-resource languages. There are unsupervised methods proposed, with small seed dictionaries [21, 75] or with no parallel data at all [22]. Some methods, however, may require large Wikipedia-similar corpora [76].

Pretrained language models, which are trained with multi-lingual data show empirical evidences that they provide cross-lingual alignment without any supervision or parallel data [77, 78]. Models are able to map embeddings of different languages to the same space. It is discussed that language similarity is important for transferring. Turkish is a agglutinative language and it is quite different than

English in terms of grammar and words. Other Turkic languages which transfer from might have been more successful, however, have even less data than Turkish. As a result, transferring from the richest language data-wise makes sense.

Cross-lingual abilities of multi-lingual BERT or mBERT are proven by many studies [14, 77, 78]. XLM-R is a network trained for learning cross-lingual representations and it significantly outperforms BERT on cross-lingual tasks [24]. As a result, we use both for out zero-shot cross-lingual classifiation task.

4.3 Datasets

4.3.1 Turkish Datasets

All the Turkish datasets used in this research are for classification task. They are either sentiment analysis datasets or to detect offense, bullying or stance. All of them, except the GSM dataset, has two labels. Table 4.1 shows the datasets with the number of data points, ratio of test set, mean length of the reviews or text in terms of words and label distributions. All the dataset are divided into test and train sets in a stratified way to have same ratio of labels. Almost all of them were already divided by the providers.

First dataset is a sentiment analysis dataset reflecting the opinions of customers of a GSM operator that offers services in Turkey. There are 17,289 tweets in the telecom dataset which are collected by using the name of the operator. All the tweets are annotated by hand with three labels: positive, negative, neutral with ratios 26.5%, 39.8% and 33.7% [79].

Second dataset is HumirSent of Hacettepe Multimedia Information Retrieval Laboratory. It includes hotel and movie reviews (from Beyazperde.com and otelpuan.com). Movie reviews dataset has 26,700 negative and 26,700 positive reviews. Hotel reviews dataset has 5,800 negative and 5,800 positive reviews. Humirsent is composed of the two review sets. [80]

Table 4.1: Turkish datasets and their statistics.

Dataset	Total	Test ratio	Avg.length	Label 0	Label 1	Label 2
GSM	17,289	0.20	11.72	6,888	4,579	5,822
Humirsent	65,000	0.50	41.00	32,500	32,500	-
Offense Det.	34,792	0.10	14.93	28,035	6,757	-
Cyberbullying	3,001	0.33	9.54	1,498	1,503	-
Books	1,400	0.33	27.31	700	700	-
Dvd	1,400	0.33	26.08	700	700	-
Electronics	1,400	0.33	37.51	700	700	-
Kitchen	1,400	0.33	32.66	700	700	-
Products All	5,600	0.33	30.89	2,800	2,800	-
Movie	10,658	0.33	15.95	5,329	5,329	-
Stance Det.	1,062	0.33	9.30	526	536	-

Third one is on offensive language detection dataset. It contains 36,232 tweets with approximately 19% of them contain some type of offensive language. Labels are further subcategorized based on the target of the offense however we only use the dataset to detect offense regardless of target [81]. Similarly, cyberbulling dataset is for cyberbullying detection, which includes 3,000 tweets with 1,500 of them includes cyberbullying [82].

Other group of datasets are consisted of movie and product reviews. Turkish movie review dataset (from Beyazperde web page) includes 5,331 positive and 5,331 negative sentences. Turkish multi domain product reviews (from Hepsiburada.com) contains reviews from different domains, for books, dvds, electronics, and kitchen appliances similar to English product reviews dataset. It has 700 positive and 700 negative reviews for each of the four categories [83]. As a new dataset, we concatenated the four product reviews datasets, to investigate the transfer among the domains.

Lastly, a Turkish stance detection dataset is included, which is formed by the tweets about 2 popular sports clubs in Turkey and annotated by the stance of the owner [84]. Although it has 2 targets, we did not consider targets in this study and only tried to detect stance regardless of the target.

4.3.2 English Datasets

Except for GSM and stance detection datasets, English datasets are chosen from the same domain with the Turkish ones in terms of subjects and classification task.

GSM dataset consists of the opinions of customers about a GSM operator and we picked a product reviews dataset as a match for it Amazon reviews [85]. Although the domain is quite different, we believe, some transfer can be possible. For stance detection dataset, we picked the SemEval-2016 English stance detection dataset, which has different domains itself [86]. We include stance detection datasets to observe cross-domain transfer.

The matching dataset for Turkish product reviews is Amazon product reviews for four different product types: books, DVDs, electronics and kitchen appliances. Each domain of reviews consists of 1,000 positive and 1,000 negative examples [87].

For Turkish movie reviews dataset, English movie reviews (from rotten tomatoes) is used with positive and negative labels. There are equal number of positive and negative samples, each of which is 5,331 [88].

As the equivalent of Humirsent dataset, we use a portion of hotel reviews dataset (from Booking.com). This dataset contains 515,000 customer reviews and scoring of 1,493 luxury hotels across Europe [89] with positive and negative labels. We combine this with movie review dataset by randomly choosing from hotel reviews to complete the number of data points to the number of test data, regarding equal ratio of labels.

For the offense detection dataset, we pick the Offensive Language Identification Dataset (OLID) dataset provided to the OffensEval 2019 participants. It contains 8,840 not offensive and 4,400 offensive tweets [90]. We also use OLID for cyberbullying dataset as well.

4.4 Models and Methods

As models, we use multi-lingual BERT base cased, mBERT and XLM-R base [24], for multi-lingual networks. Both networks are trained with multi-lingual data including Turkish and English. For cross-lingual transfer, we finetune the models with English equivalents of datasets and test the performance of Turkish test sets. We only use test sets for comparison with the monolingual networks.

For comparison purposes, we use multi-lingual BERT base cased, BERTurk base cased and DistilBERTurk base cased [91] models to evaluate monolingual performance of Turkish datasets. BERTurk and DistilBERTurk are pretrained with Turkish data only and they do not have classification layer. We add a classification layer to both, similar to other networks provided for sequence classification.

4.5 Experimental Setup

For monolingual experiments, we simply finetune the networks with 3 epochs, and test for performance with test sets. We use the provided train and test sets for Turkish datasets.

For cross-lingual experiments, we finetune the multi-lingual netoworks with English datasets. When the English dataset is relatively larger or ratios of the classes are different, we choose randomly to decrease the data size to the Turkish datasets' size or to make the class ratios same. Classes of Turkish datasets are usually 1 or 0, indicating the existance of a stance, sentiment or offensive language. When English datasets have higher number of classes, like rates or stars form 1 to 5, we use 1 and 2 stars for negative sentiment, 4 and 5 for positive sentiment and if there is neutral sentiment in the equivalent Turkish dataset, 3 for neutral.

For bilingual experiments, we use the same training set of English datasets

from the crosslingual experiments; however, this time we add a small amount of Turkish data from training sets. During the experiments, we concluded that the best and most reliable results are seen when we add Turkish data randomly to the English data. We test for adding one data from each class, adding 10 data points, 100 points and 500 points each has equal ratio of classes.

For both monolingual and bilingual experiments, we use the same parameters whenever possible. For finetuning the networks, we use Adam Optimizer [92] with learning rate 10^{-5} and epsilon 10^{-8} for each experiment. Since sequence length can be different among cross datasets and considering in most cases there are outliers which are quite long, we use the 95th percentile of each sequence lengths of each dataset as maximum length parameter. We share the maximum lengths and batch sizes for each experiment in the results section of the chapter.

As performance measures, we use accuracy and F1 score. In many of the datasets, the classes are distributed equally so those two measures are close for the most experiments.

4.6 Experimental Results and Discussion

To have baseline results and to see the performance of newly trained BERTurk and DistilBERTurk networks, we first fine-tune them and test them with Turkish data. Table 4.2 shows the results for the experiments. Batch size and maximum sequence lengths used for the datasets are given in the all tables. All other parameters of the experiments in this section of the chapter are used as mentioned in the Experimental Setup section of the chapter.

To compare with cross-lingual and bi-lingual finetuning results, we train and test mBERT and XLM-R with mono-lingual (Turkish) data. Table 4.3 shows the results for the experiments.

After experiments for baselines, we do cross-lingual finetuning experiments.

Table 4.2: Classification results with monolingual networks.

Dataset	Batch size	Max length	BERTurk Accuracy	BERTurk F1score	DistilBERTurk Accuracy	DistilBERTurk F1score
GSM	128	35	0.770	0.770	0.709	0.710
Humirsent	32	123	0.929	0.922	0.913	0.913
Offense Det.	32	35	0.877	0.870	0.872	0.866
Cyberbullying	32	18	0.923	0.923	0.890	0.890
Books	32	58	0.887	0.887	0.835	0.835
Dvd	32	59	0.827	0.827	0.773	0.772
Electronics	32	76	0.879	0.879	0.831	0.831
Kitchen	32	67	0.788	0.788	0.740	0.740
Products All	32	66	0.853	0.853	0.832	0.832
Movie	32	36	0.887	0.887	0.862	0.862
Stance Det.	32	18	0.869	0.868	0.826	0.826

We train mBERT and XLM-R with English datasets and test with Turkish data from the similar domain. Table 4.4 shows the results for the experiments.

Lastly, for the bi-lingual finetuning experiments, we mix English and Turkish training data, by adding a limited amount of data points from Turkish, to test for the challenge of low resource cases. We add 1 data point for each class, and 10, 100 and 500 data points which has equal number of classes. Table 4.5 shows the results of the experiments.

Table 4.6 shows the average percentage changes of accuracies of the experiments in rows according to experiments in the columns. Table clearly shows that monolingual experiments give the higher results and XLM-R outperforms mBERT in almost every case. Adding some amount of target language data to the training set, also improves accuracy compared to the fully cross-lingual experiments.

Results show that both mono-lingual networks, BERTurk and DistilBERTurk outperforms multi-lingual networks, as expected. We see that, for Turkish, multi-lingual networks are not as successful as mono-lingual networks. We see that most of the languages in the training data of the multi-lingual networks are distant

Table 4.3: Classification results with multi-lingual networks.

Dataset	Batch size	Max length	MultiBERT Accuracy	MultiBERT F1−score	XLM-R Accuracy	XLM-R F1−score
GSM	128	35	0.647	0.641	0.695	0.690
Humirsent	32	123	0.908	0.908	0.937	0.937
Offense Det.	32	35	0.847	0.831	0.853	0.841
Cyberbullying	32	18	0.765	0.765	0.820	0.820
Books	32	58	0.801	0.800	0.853	0.852
Dvd	32	59	0.729	0.727	0.794	0.793
Electronics	32	76	0.788	0.787	0.835	0.835
Kitchen	32	67	0.688	0.688	0.714	0.714
Products All	32	66	0.775	0.775	0.823	0.823
Movie	32	36	0.824	0.824	0.837	0.837
Stance Det.	32	18	0.741	0.740	0.695	0.690

languages to Turkish, as a result multi-lingual networks are still not competitive enough for Turkish. When we compare performances of mBERT and XLM-R with all Turkish data, however, we see that XLM-R highly outperforms mBERT and shows a closer performance to BERTurk.

Cross-lingual experiments show performances over the majority classification for the most datasets. In the experiments with GSM and stance detection datasets, results are almost equal to the majority classification rates, so those experiments show no evidence of transferring. English equivalents of those datasets are from different domains and we see that domain similarity is important for transferring from English to Turkish. Again, XLM-R shows better results, however results are not competitive in any way with the mono-lingual finetuning results.

Table 4.4: Classification results with crosslingual finetuning.

Train (Eng)	Test (Tr)	Batch size	Max. length	MultiBERT Acc.	MultiBERT F−score	XLM-R Acc.	XLM-R F−score
Prod. Review	GSM	16	142	0.407	0.375	0.492	0.415
Hotel+Movie	Humirsent	32	101	0.650	0.637	0.873	0.872
OLID	Offense Det.	32	46	0.765	0.732	0.804	0.780
OLID	Cyberbullying	32	47	0.589	0.540	0.631	0.588
Books	Books	16	200	0.569	0.569	0.708	0.694
Dvd	Dvd	16	200	0.576	0.573	0.688	0.671
Electronics	Electronics	16	200	0.604	0.602	0.701	0.676
Kitchen	Kitchen	16	200	0.541	0.527	0.658	0.637
Products All	Products All	16	200	0.554	0.550	0.695	0.672
Movie	Movie	32	37	0.698	0.607	0.778	0.777
Stance Det.	Stance Det.	32	25	0.456	0.376	0.464	0.385

Bi-lingual finetuning experiments show similar results to cross-lingual finetuning experiments. We see that XLM-R, again, give better results. Adding really small amounts of Turkish to training data do not have a major effect on the accuracy. However adding 100 data points or 500 data points improves accuracy significanly for smaller datasets. For the small datasets, 500 data points is nearly half of the training set but still results with 500 points added are below the monolingual experiments.

Table 4.5: Classification results with bilingual finetuning.

Train (Eng)	Test (Tr)	Batch size	Max. length	MultiBERT / 1 Acc.	MultiBERT / 1 F−score	XLM-R / 1 Acc.	XLM-R / 1 F−score	MultiBERT / 10 Acc.	MultiBERT / 10 F−score	XLM-R / 10 Acc.	XLM-R / 10 F−score
Prod.Reviews	GSM	16	140	0.419	0.380	0.490	0.425	0.407	0.394	0.500	0.442
Hotel+movie	Humirsent	32	101	0.648	0.646	0.868	0.866	0.649	0.646	0.862	0.862
OLID	Offense Det.	32	46	0.783	0.732	0.800	0.781	0.738	0.736	0.805	0.796
OLID	Cyberbullying	32	47	0.537	0.428	0.649	0.618	0.563	0.501	0.711	0.701
Books	Books	16	200	0.526	0.477	0.710	0.710	0.623	0.620	0.740	0.739
Dvd	Dvd	16	200	0.568	0.556	0.675	0.653	0.517	0.516	0.712	0.711
Electronics	Electronics	16	200	0.561	0.545	0.714	0.692	0.571	0.567	0.729	0.711
Kitchen	Kitchen	16	200	0.489	0.450	0.654	0.623	0.535	0.497	0.654	0.647
Products All	Products All	16	200	0.540	0.505	0.700	0.678	0.563	0.553	0.706	0.687
Movie	Movie	32	37	0.602	0.592	0.776	0.774	0.601	0.594	0.783	0.783
Stance Det.	Stance Det.	32	25	0.496	0.487	0.538	0.536	0.467	0.461	0.604	0.591

Table 4.5: Classification results with bilingual finetuning.

Train (Eng)	Test (Tr)	Batch size	Max. length	MultiBERT / 100 Acc.	MultiBERT / 100 F−score	XLM-R/ 100 Acc.	XLM-R/ 100 F−score	MultiBERT / 500 Acc.	MultiBERT / 500 F−score	XLM-R / 500 Acc.	XLM-R / 500 F−score
Prod.Reviews	GSM	16	140	0.382	0.377	0.521	0.514	0.455	0.442	0.553	0.550
Hotel+movie	Humirsent	32	101	0.716	0.716	0.873	0.873	0.785	0.784	0.889	0.889
OLID	Offense Det.	32	46	0.618	0.655	0.727	0.749	0.644	0.678	0.716	0.742
OLID	Cyberbullying	32	47	0.696	0.696	0.800	0.800	0.781	0.780	0.841	0.840
Books	Books	16	200	0.699	0.700	0.751	0.749	0.810	0.809	0.786	0.786
Dvd	Dvd	16	200	0.636	0.635	0.745	0.744	0.751	0.751	0.784	0.783
Electronics	Electronics	16	200	0.623	0.636	0.745	0.733	0.773	0.773	0.779	0.772
Kitchen	Kitchen	16	200	0.576	0.576	0.669	0.653	0.667	0.666	0.712	0.707
Products All	Products All	16	200	0.640	0.638	0.729	0.719	0.698	0.697	0.758	0.753
Movie	Movie	32	37	0.621	0.621	0.787	0.787	0.685	0.685	0.792	0.792
Stance Det.	Stance Det.	32	25	0.561	0.491	0.610	0.604	0.741	0.740	0.698	0.697

Table 4.6: Percentage change in the performances of experiments in the rows comparing to the ones in the columns.

	BERTurk	dBERTurk	mBERT	XLM-R	cr-mBERT	cr-XLM-R
bi-XLM-R/500	−13	−9	−2	−6	32	13
bi-XLM-R/100	−16	−13	−7	−10	25	8
bi-XLM-R/10	−18	−14	−9	−12	23	5
bi-XLM-R/2	−20	−17	−11	−15	19	2
bi-mBERT/500	−18	−14	−8	−12	24	6
bi-mBERT/100	−29	−26	−21	−24	7	−8
bi-mBERT/10	−34	−32	−27	−30	−2	−16
bi-mBERT/1	−35	−32	−28	−31	−3	−17
cr-XLM-R	−21	−18	−12	−16	17	
cr-mBERT	−33	−30	−25	−28		
XLM-R	−7	−2	4			
mBERT	−10	−6				
distilBERTurk	−4					

Table 4.6: Precentage change in the performances of experiments in the rows comparing to the ones in the columns.

	bi-mBERT/1	bi-mBERT/10	bi-mBERT/100	bi-mBERT/500	bi-XLM-R/2	bi-XLM-R/10	bi-XLM-R/100
bi-XLM-R/500	37	35	24	7	11	7	5
bi-XLM-R/100	30	29	18	3	6	2	
bi-XLM-R/10	27	26	16	1	3		
bi-XLM-R/2	23	22	13	−2			
bi-mBERT/500	28	26	15				
bi-mBERT/100	11	9					
bi-mBERT/10	1						

4.7 Conclusion and Future Work

Zero-shot cross-lingual transfer learning is still not competitive in terms of its performance for transferring between distant languages. On the other hand, for large datasets coming from similar domains, cross-lingual experiments give satisfactory results. Even though cross-lingual transferring from English to Turkish would not allow for an analysis without Turkish training data yet, it can help enriching the data at hand, or may help in data creation and annotation tasks. As a future work, cross-lingual transfer can be used as a starting point for an iterative data labeling process, using the labeled data with bilingual finetuning and iterating the process until it reach to an equilibrium performance.

Chapter 5

Stylometric Time-based Analysis of Ahmet Hamdi Tanpınar's Works

5.1 Introduction

Texts can be analyzed in many ways such as text categorization, text clustering, concept/entity extraction, document summarization, sentiment analysis, etc. The purpose to analyze a text can be also very different like for security or academic reasons or for business and marketing related reasons, etc. Stylometry is also a method to analyze a text and it is defined as a pursuit to capture the characteristics of the style of a particular text by a variety of quantitative criteria, usually lexical [93].

In stylometry, researchers analyze writing styles of authors using objective measures. Various style markers (measurable attributes) are created for this aim and pattern of those measures in the text of interest are tested using statistical methods. Anticipated patterns are used to answer stylometric problems. Such problems can be illustrated as authorship attribution (i.e., assigning author to

work) and stylochronometry (i.e., assigning date to work). In stylometry, there is a supervised classification problem, since a set of pre-classified works exist, and the aim is to classify a new work into these existing classes. In terms of the problem and the aim, stylometry is similar to data mining and clustering [8].

In this study, we analyze the different types of works of Ahmet Hamdi Tanpınar. We will use six style measures: "sentence length in terms of words", "most frequent words", "word length of type", "word length of token", "syllable count of type" and "syllable count of token". We study the differences among the works with all the style markers and examine change in word length and most frequent words with time as in [7, 94].

5.1.1 Motivation and Importance

Such studies in literature are conducted to author assignments or date assignments in general. They help to assign authors or date to the newly found works with unknown writers or date by analyzing suspected writers' works and the work newly found. The results are compared and if it is proven statistically that the found work's style is significantly similar to a writer's works or works written in a period of time, then it can be assigned to that writer or time period. Of course, such studies can also serve as means to help literary critics and an intellectual tool. They can support or prove wrong some critics' claims by statistical observations. Thus, stylometric studies are important both in terms of literature and from an intellectual point of view.

In this chapter, we aim to make a stylometric analysis of Ahmet Hamdi Tanpınar's works. He is one of the most known and important writers of Turkish literature. Many literature critics in Turkey consider him as the most important novel writer in Turkish language due to his different style of writing. Also, he wrote different types of works such as poems, articles, essays and stories during his writing period. He was influenced by different literature trends from Europe as well as one can see the marks of French language in his works [95].

Ahmet Hamdi Tanpınar's works are not stylometricly analyzed before even if he is considered as the one of the greatest writers of Turkish. This study is important in terms of such study will be a first for Ahmet Hamdi Tanpınar, moreover the method can be applied easily for other writer's works as well.

5.2 Related Work

Even though such stylometric analyses are done by for different language data, there are not many for Turkish language. Can and Patton have several studies examining change of writing style In Turkish. In one of their studies, they study two writers over their writing period [8]. In another, they investigate change of word characteristics of Turkish over a century by analyzing 40 different writers [7]. Similar to these studies, they analyze Yaşar Kemal's works, also a well-known writer in Turkish literature, with a method similar to the one in this section [94].

5.3 Experimental Environment And Design

5.3.1 Test Data

For the study, first we have collected works of Ahmet Hamdi Tanpınar. Table 5.1 shows works of him that are analyzed in this study with the first publication year.

Abdullah Efendinin Rüyaları was first published in 1943 and *XIX. Asır Türk Edebiyatı Tarihi* is three volume study, each volume first published in 1946, 1966 and 1967 respectively. Works without the first publication year are the ones Ahmet Hamdi Tanpınar wrote along with his writing period and they are published after his death. So, those works do not have a particular time that they can be representatives of, and we can say that they reflect the style of Ahmet Hamdi Tanpınar in general including works from all his writing period. *Bütün Şiirleri*

can also be included to his works which are published after his death since it is published a year before his death and includes all his poets from different periods.

Table 5.1: Work titles and first publication years.

Title of the Work	First Publication Year
Abdullah Efendinin Rüyaları	1943
Mahur Beste	1944
Beş Şehir	1946
Huzur	1948
Sahnenin Dışındakiler	1950
Saatleri Ayarlama Enstitüsü	1954
Yaz Yağmuru	1955
XIX. Asır Türk Edebiyatı Tarihi	1949 - 1966 - 1967
Aydaki Kadın	1962
Yahya Kemal	1962
Bütün Şiirleri	1961
Edebiyat Üzerine Makaleler	1969
Yaşadığım Gibi	1970
Mektuplar	1974
Mücevherlerin Sırrı: Derlenmemiş Yazılar, Anketler ve Röportajlar	2002

5.3.2 Experimental Design

For the experiment, firstly we analyze each work of Ahmet Hamdi Tanpınar separately and compare the results to make a comment about his style in general. In the light of the results, some further analysis is aimed such as regression analysis to monitor differences in the works by their publication year.

5.3.2.1 Selection of Block Size and Style Measures

Block size should be a value that enable us to have large samples to apply statistical tests. In order to decide on it, we also consider total sizes of works and pick the block size as 5000 words since it is both large enough to represent the characteristics of a work and allow us have enough numbers of blocks.

As indicate before, we will use “sentence length in terms of words”, “most frequent words”, “word length of type”, “word length of token”, “syllable count of type” and “syllable count of token” as style measures.

5.4 Experimental Results

In this section, firstly style measures mentioned before are calculated for each work. Most frequent words for all of the works are also founded to grasp the style of the writer in general. Using the measures calculated, principle component analysis (PCA) is applied especially for visualization purposes. This is followed by regression analysis to see the relation between average token length sand the publication years of the works. Similar analysis is done using the same size blocks of only novels and only letters.

5.4.1 Style Marker of Works

Each style marker for each work is calculated separately and they are used to make further analyses. Table 5.2 shows the style measures for each work. When they are examined in general, we see that Ahmet Hamdi Tanpınar uses longer sentences in his essays and articles while he uses shorter sentences in his novels. However, the shortest average sentence length belongs to his letters, *Mektuplar*, so we can say that in his daily life he did not use long sentences as in the other works. Also, his letters and poems have shorter average token and type lengths compared to other works, so it can be deducted that he uses a simple language in both his poems and daily language, compared to his other works.

In Table 5.3, we see most frequent 10 words for each of the works. The aim of examining them for each work separately is to see if they are changing work by work or by the type of work. Mostly, they seem to be similar, however we see that for his novels and stories, character names are among the most frequent words as well as the word ‘bey’, which a formal word in Turkish meaning Mr. As a

Table 5.2: Style measures of each work.

Name of the work	Number of tokens	Number of types	Average token length	Average type length	Average sentence length	Average syllabus length token	Average syllabus length type
Abdullah Efendinin Rüyaları	10,860	4,397	6.036	7.691	12.976	2.570	3.260
Aydaki Kadın	71,654	18,396	6.020	8.276	7.526	2.563	3.494
Beş Şehir	35,253	12,900	6.232	8.109	14.717	2.633	3.392
Edebiyat Üzerine Makaleler	165,584	34,326	6.166	8.612	14.283	2.594	3.565
Huzur	98,665	23,507	6.147	8.518	10.275	2.642	3.630
Mahur Beste	39,055	12,788	6.152	8.139	12.493	2.628	3.440
Mektuplar	62,155	18,165	5.945	7.973	7.036	2.493	3.343
Saatleri Ayarlama Enstitüsü	92,134	22,186	6.088	8.521	9.409	2.592	3.594
Sahnenin Dışındakiler	75,432	19,275	6.112	8.438	8.788	2.603	3.558
Şiirler	8,941	4,034	5.753	6.938	15.769	2.350	2.836
Yahya Kemal	39,399	12,418	6.218	8.111	14.809	2.630	3.371
Yaşadığım Gibi	86,742	24,269	6.235	8.568	13.090	2.625	3.575
Yaz Yağmuru	14,900	5,538	5.989	7.777	7.348	2.562	3.300

result, we can say that Ahmet Hamdi Tanpınar uses formal character names such as 'Mümtaz Bey' and he uses character names frequently in his novels which is leading us to the deduction that main events in his novels happens around some main characters.

Table 5.3: Most 10 frequent words for each work.

Title of the Work	Frequent words
Abdullah Efendinin Rüyaları	bir, ve, bu, abdullah, gibi, bütün, fakat, o, kadar, içinde
Aydaki Kadın	bir, ve, bu, gibi, o, de, selim, sonra, kadar, bütün
Beş Şehir	bir ve bu gibi o kadar için de ile her
Edebiyat Üzerine Makaleler	bir, ve, bu, gibi, o, için, de, kadar, bütün, çok
Huzur	bir, bu, ve, o, gibi, fakat, için, kadar, mümtaz, de
Mahur Beste	bir, bu, ve, gibi, o, için, kadar, her, bey, fakat
Mektuplar	bir, ve, bu, de, fakat, çok, ne, da, gibi, ben
Saatleri Ayarlama Enstitüsü	bir, bu, ve, o, de, gibi, için, kadar, sonra, daha
Sahnenin Dışındakiler	bir, bu, ve, o, de, gibi, kadar, fakat, için, çok
Şiirler	bir, ve, bu, gibi, her, ne, bütün, de, ben, o
Yahya Kemal	ve, bir, bu, yahya, gibi, de, o, olan, kadar, daha
Yaşadığım Gibi	bir , ve, bu, gibi, o, için, kadar, de, her, çok
Yaz Yağmuru	bir, bu, ve, gibi, o, sonra, de, için, fakat, kadar

Figure 5.1 shows the top 100 most frequent words the writer used in all of his works. Most of them are stopwords, as can be guessed. However, there are some interesting words of choice of the writer, which cannot be seen in other writers' top 100 most frequent words. 'Kadın', 'zaman' and 'gece' can be given as an example to those words whose meanings are woman, time and night [96]. He has a novel about time named *Saatleri Ayarlama Enstitüsü* which means Institute of Time Regulation, however in this novel the frequency of word time is 275. Frequency of the word among all the works is 2530. Thus, we can say that choice of word 'time' so frequently does not entirely depends on the novel and he is interested in time, woman and nights in general.

Figure 5.1: Most frequent 100 words listed in decreasing order of occurrence frequencies.

bir, ve, bu, gibi, o, de, kadar, için, fakat, çok, her, bütün, da, daha, ne, sonra, ki, şey, onun, ile, hiç, büyük, zaman, kendi, olan, en, ben, içinde, onu, başka, diye, iki, biraz, küçük, bile, beraber, böyle, eski, vardı, bana, olduğu, güzel, bey, belki, var, birdenbire, sadece, değil, ona, beni, iyi, ilk, yeni, gün, şimdi, idi, veya, hemen, benim, son, doğru, nasıl, daima, tek, birkaç, insan, yahut, kendisini, arasında, genç, kendisine, türlü, asıl, bunu, yalnız, hattâ, evvel, çünkü, yavaş, eden, mi, hiçbir, olduğunu, yığın, kadın, tekrar, hep, oldu, olarak, aynı, defa, yine, dedi, artık, tam, gece, ama, yahya, vardır, sanki

5.4.2 Clustering of Works Using Style Measures

The PCA plot, Figure 5.2, provides of the similarities among the works by using the style measures calculated except most frequent words. As can be seen, the same type of works appears closer to each other. His novels and stories are gathered at one side of the plot while his essays and articles were grouped at the other side. We see that *Mahur Beste* and *Abdullah Efendinin Rüyaları* are somewhat separated from other novels and stories and this is probably due to their publication time being earlier than the other novels. This deduction might lead to the assumption that his writing style changed over time even though further analysis would be needed to say so.

Poems also seem to be separated from other works, and this is something expected since it has quite different style measure due to being a completely different work type from the others.

5.4.3 Regression Analysis Using Style Measures

A linear regression analysis is done by using the style measure average token length and the publication year of the work. The aim of this analysis to discover a relationship between years and token length, i.e. to show that as years passes, average token length become longer as is the case for other writers in Turkish literature lived at the same time period with Ahmet Hamdi Tanpınar [7].

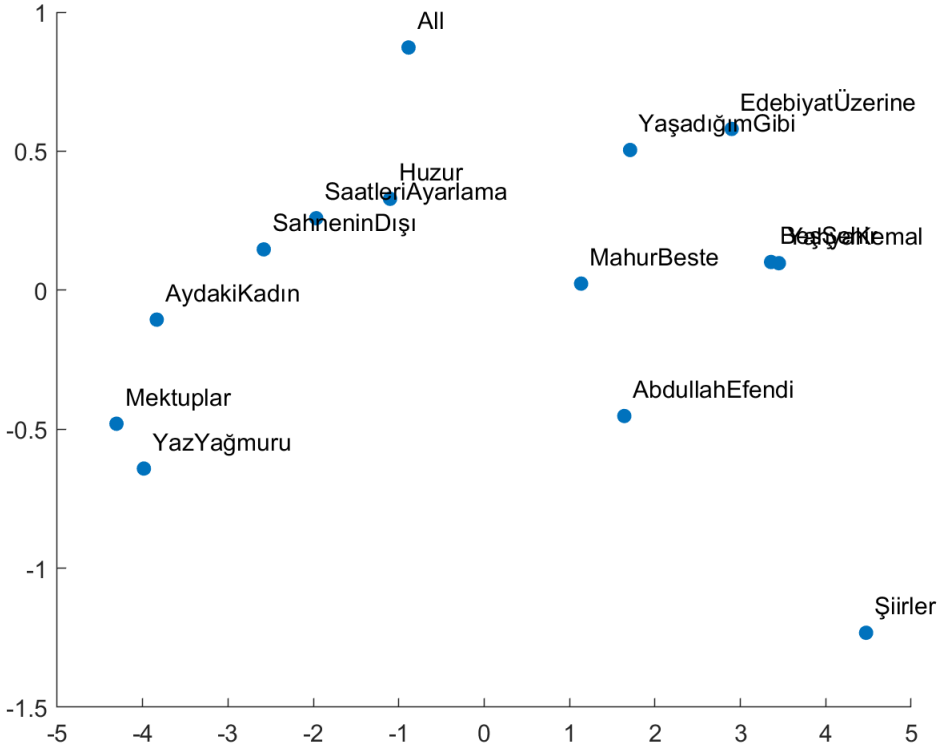


Figure 5.2: PCA plot of all the works with style markers.

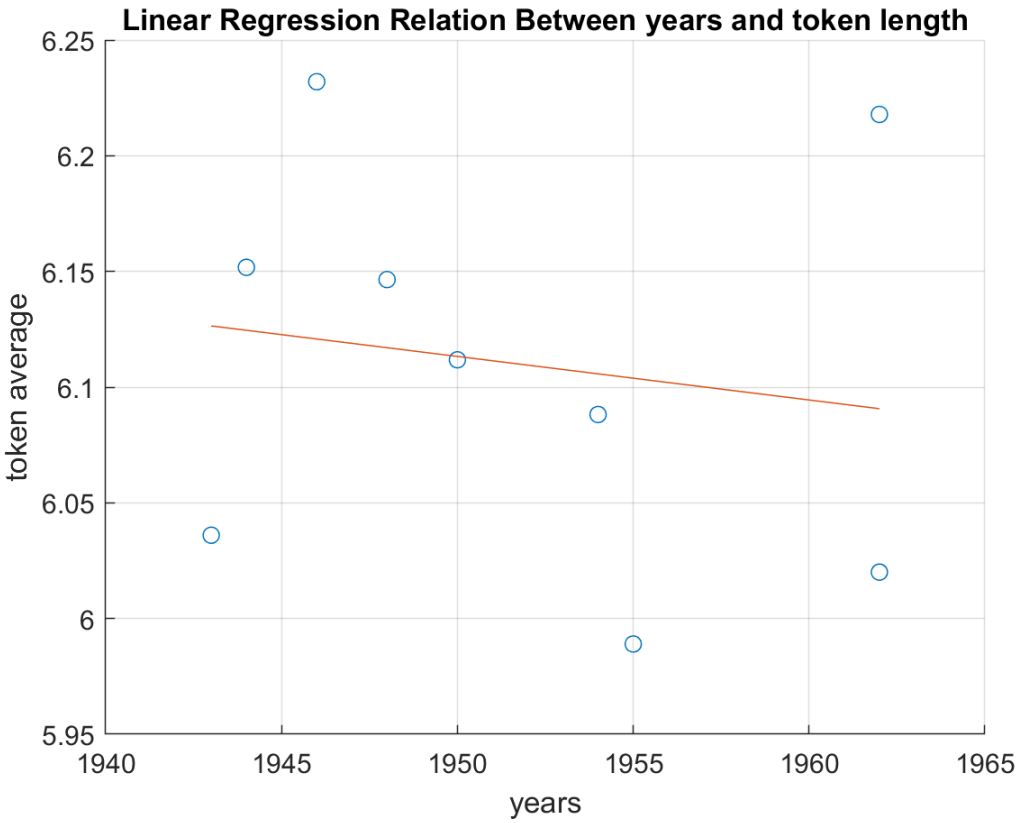


Figure 5.3: Linear Regression plot using average token length and first publication year of the works.

Figure 5.3 shows the linear regression plot using the works whose publication years are certain, i.e. publication year is known certainly and it is not after the death of Ahmet Hamdi Tanpınar. Works fit to this definition are *Abdullah Efendinin Rüyaları, Mahur Beste, Beş Şehir, Huzur, Sahnenin Dışındakiler, Saatleri Ayarlama Enstitüsü, Yaz Yağmuru, Aydaki Kadın, Yahya Kemal* and *Bütün Şiirleri*. Bütün Şiirler or Poems is published in 1961 even though poems in it are written in different years. After applying regression with those works, it is seen that Şiirler or Poems affect the trend a lot since it has a short average token length. So, it causes the relation between two values to have sharp slope, which is not natural since poems are different than the other works and it should not be evaluated with others. Then it is taken out from the data and regression was fitted again and the result was as in Figure 5.3. The prediction equation was given by equation 5.1.

$$AvgTokenLength = 9.7833 - 0.0019 * PublicationYear \tag{5.1}$$

So, as the years pass, average token length increases since coefficient of the publication year is negative, i.e. two values have negative relationship. This result is quite surprising since in the paper Change of Word Characteristic in Turkish, authors made a similar regression analysis for 40 different novels from 40 different writers of last century and they showed that average token length increases as the publication year of a work increases [7]. They also showed that in another paper, for two writers' works Çetin Altan and Yaşar Kemal, as the age of work increases, average token length increases [8]. As a result, Ahmet Hamdi Tanpınar's style shows a sharp difference than his contemporaries.

The regression is applied for different types of works of him. This might cause the decreasing slope since different types of works might(would) have different styles and quite different average token length, causing such slope due to novels being usually published in early writing period of his and essay and articles being published later.

We thought that it might be more meaningful to apply this analysis to the

same type of works of him separately to see the relations. However, he has only 5 novels and 2 story books, 4 essay and article books in total and we have only published texts of his letters. So, number of works are not enough to apply the regression directly to the measure of each work. This led to us splitting the works into same sized word blocks, which have the same number of words from each work. It was easy to label those blocks for publication year for novels and stories since they have the same label as the whole work. However it was not that easy for essays, articles and letters since this works are collection of different writings of him each written at different times.

For letters, we have separated letters written in 1930s, 1940s, 1950s and 1960s and formed equal sized blocks using this separated letters to be able to label them easier. However, keep in mind that there relatively more letters written in 1950s and 1960s than 1930s and 1940s so the difference of the number of blocks may be misleading as well.

For articles and essays, we have not done any separation and use the works as they are to keep each of them as the average work of all his works. In this case, out expectation was not to have a significant change in the token length with years since each of those essays and articles include works from almost each year of the writer's writing period.

Figure 5.4 shows the regression analysis for his novels and stories by splitting them into same size word blocks of 5000 words. In total 81 blocks were formed and 7 different works and number of blocks for each can be seen in the figure. Blocks of a particular work was spread over the publication year of the work so there are horizontal patterns in the figure. As can be seen, average token length decreases also for the novels and stories of the writer. The prediction equation was given by equation 5.2.

$$AvgTokenLength = 20.5046 - 0.0074 * PublicationYear \tag{5.2}$$

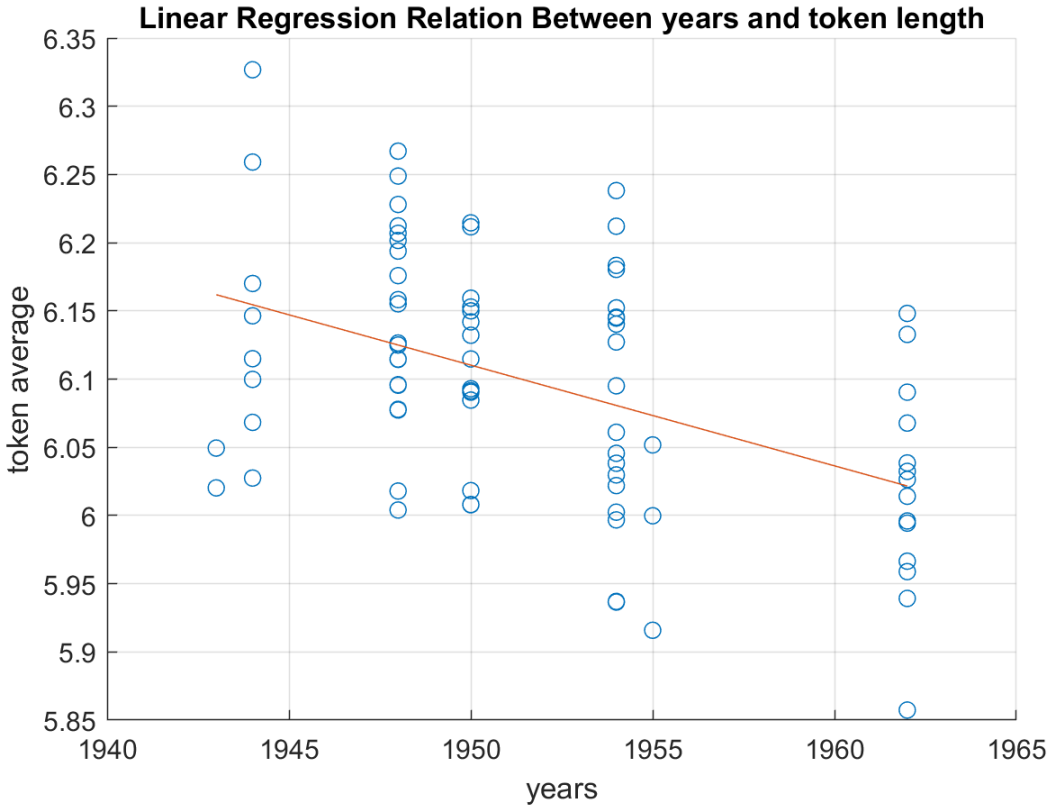


Figure 5.4: Linear Regression plot using average token length and first publication year of the novels and stories only.

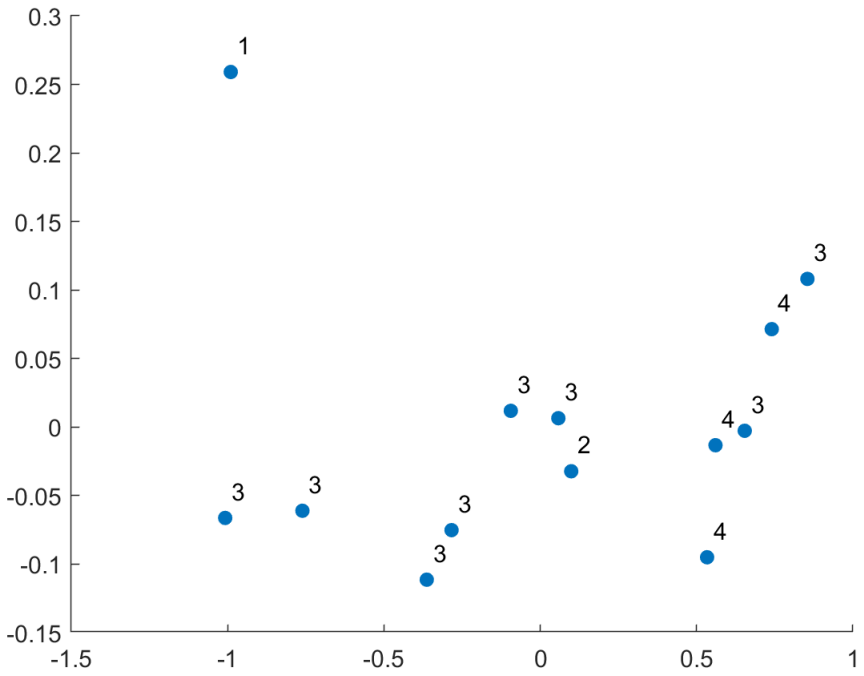


Figure 5.5: PCA Plot of word blocks of letters of size 5000.

The slope of the negative relation between average token length and the publication year is greater in the regression line fitted for his novels and stories than the regression line fitted for all his works. So, we can assume that for other work types we might see a positive relation between them.

Figure 5.5, PCA plot above shows the word blocks of letters of size 5000 and their labels indicates the year of the written date on the letters. 1 stands for 1930s, 2 for 1940s, 3 for 1950s and finally 4 stands for 1960s. We had a total of 13 blocks and as can be seen, there are only 2 blocks of letters which was written in 1930s and 1940s. However we see that 1950s and 1960s seems to be separated from them although they do not have a certain separation from each other.

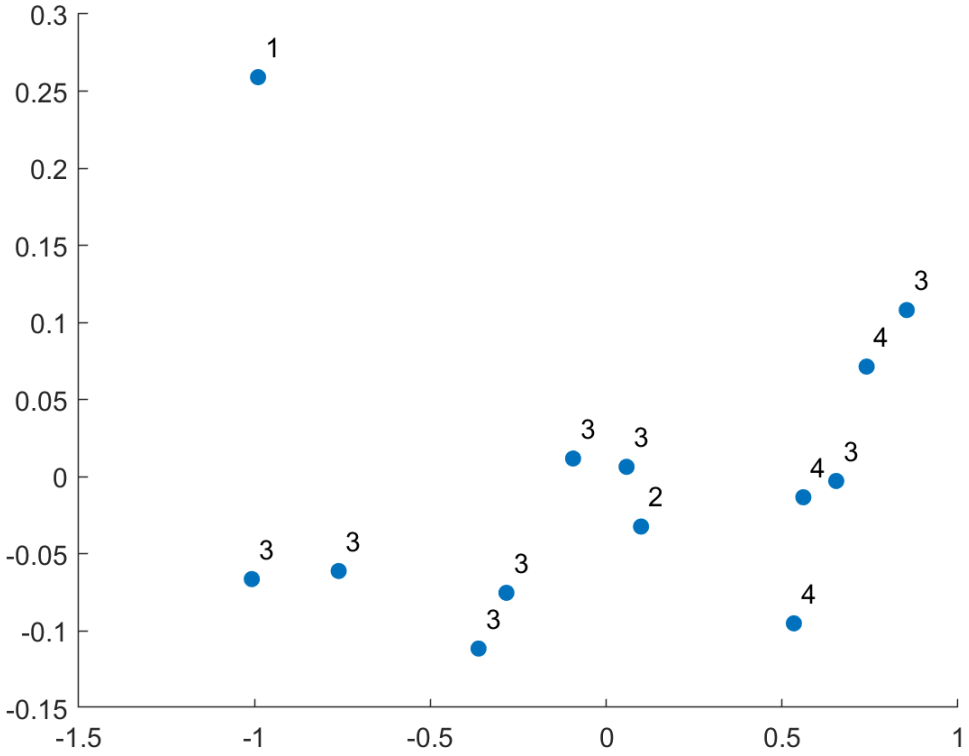


Figure 5.6: Linear Regression plot using average token length and date of the letters only (with block size 5000).

Linear regression analysis was applied for letters as well and Figure 5.6 was the result. The prediction equation is given by equation 5.3.

$$AvgTokenLength = -3.1667 + 0.0047 * PublicationYear \tag{5.3}$$

For letters, we see that average token length increases for Ahmet Hamdi Tanpınar as years pass and this is also a surprising result since his works in total have a negative relation between average token length and the publication year of the work. Here we see that for his daily language, Ahmet Hamdi Tanpınar's style shows the same pattern with his colleagues. Hence we can make the interpretation that Ahmet Hamdi Tanpınar reviews his works many times and even after years they are first written that the natural pattern of decreasing token average is not seen in his novels and stories. Of course, this claim needs to be further investigated.

It is mentioned before that relatively more letters written in 1950s and 1960s than 1930s and 1940s so the difference of the number of blocks may be misleading. In order to tackle with this issue, we had decreased the size of the equal word blocks to 2500 to have more word block for the letter written in 1930s and 1940s. There was not a significant increase in the number of blocks for those decades although we had a total of 25 word blocks this time. The regression plot formed as in Figure 5.7, which was not really different from the first one. The prediction equation was given by equations 5.4.

$$AvgTokenLength = -3.9681 + 0.0051 * PublicationYear \tag{5.4}$$

Lastly we applied regression analysis to equal sized word blocks of the writers' essays and articles. There were some works between the essay and articles of Ahmet Hamdi Tanpınar whose number of tokens are higher than the other works relatively such as *Edebiyat Üzerine Makaleler* and *Yaşadığım Gibi*. As a result, number of word blocks whose size equals to 5000 was 65. Figure 5.8 above shows the relationship between the average token length of those word blocks and their

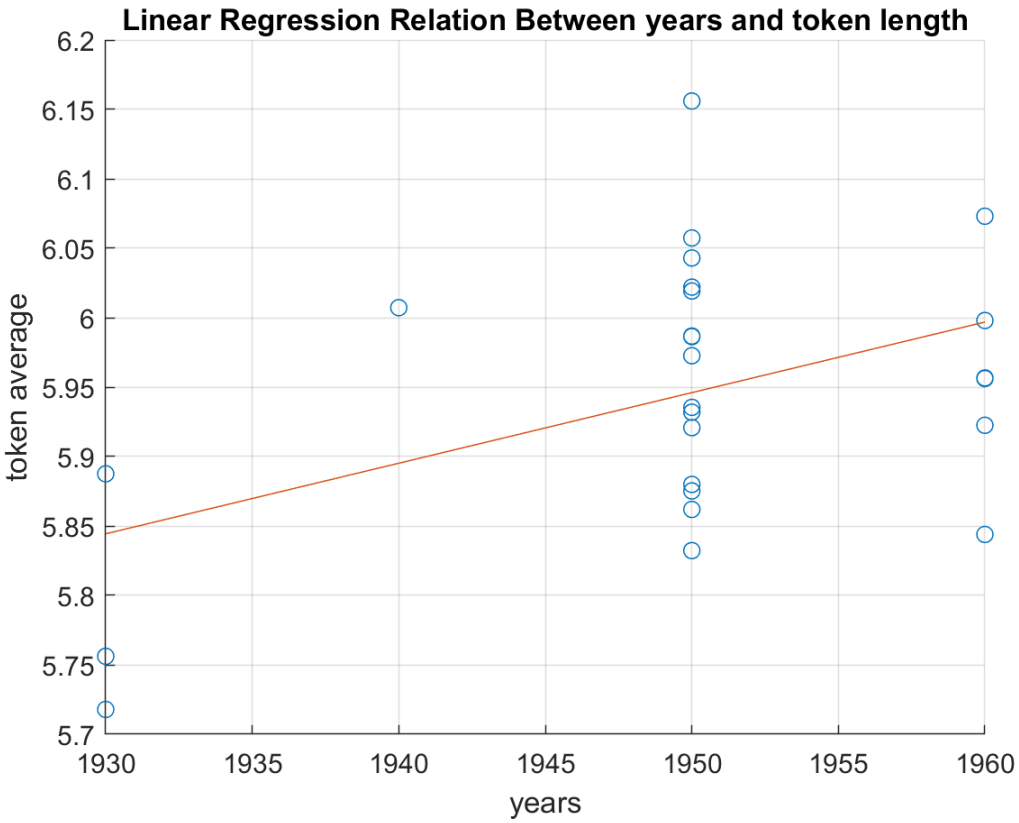


Figure 5.7: Linear Regression plot using average token length and date of the letters only (with block size 2500).

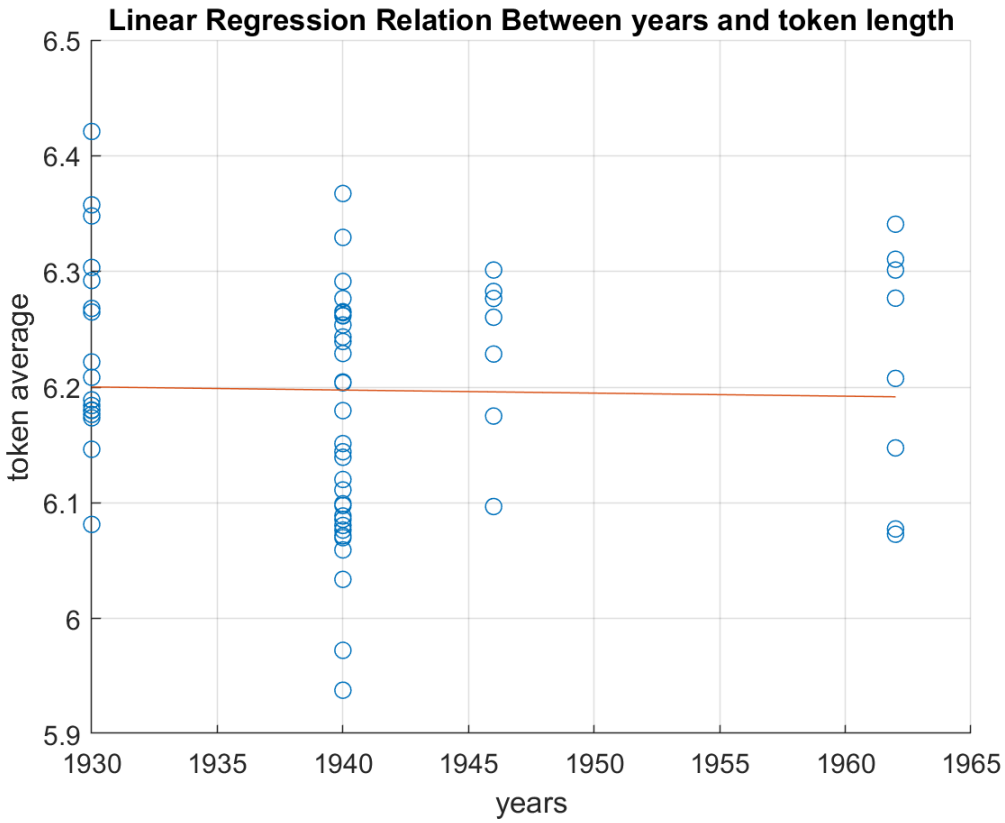


Figure 5.8: Linear Regression plot using average token length and first publication year of essays and articles only.

publication year. However, as will be remembered, those works are collections of other writings authored in different times and published in different newspapers, magazines and journals. As a result, the year of publication is not really indicative and each work can be thought of as the reflection of his all works, or mean of his all works in other words. Also, please note that for the ones whose publication year was later than the writer's death, we have put an publication year in between his writing period just for the purpose to put them on the plot. In this manner, our expectation was to see no change in the average token length or to see really small, insignificant changes as the years pass. The prediction equation is given by equation 5.5.

$$AvgTokenLength = 6.7134 + 0.0003 * PublicationYear \tag{5.5}$$

In equation 5.5, we see that publication year has a really small coefficient compared to the equations of other types of works. So, we can say that the assumption holds and those works reflect the average.

PCA plot in Figure 5.9 shows the word blocks of the essays and articles. It can also be seen here that there is not a sharp separation between the blocks, with other style measures included, so it supports the regression analysis results in Figure 5.8.

5.5 Conclusion and Future Work

Aim of the study was to make a stylometric analysis of Ahmet Hamdi Tanpınar's works. It is done by the calculation of six style measures: "sentence length in terms of words", "most frequent words", "word length of type", "word length of token", "syllable count of type" and "syllable count of token". PCA results using these style measures showed visual separation between the different types of works and between same types of works which has very different publication times. Those results led us to make a type-wise analysis of the works. After

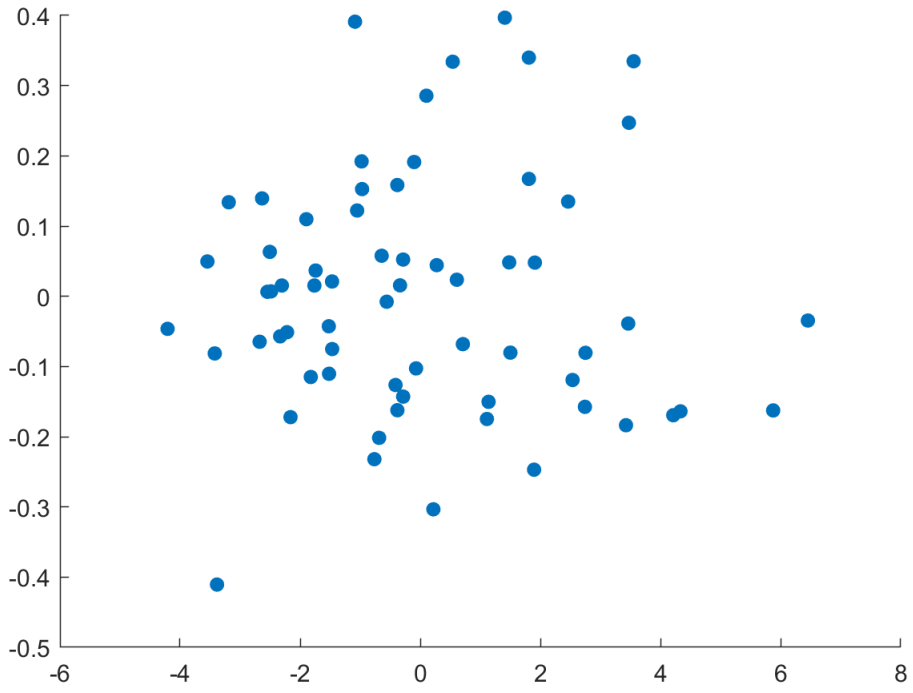


Figure 5.9: PCA Plot of word blocks of essays and articles of size 5000.

doing so, we see that the average token length relation with the publication year changes from type to type. It might be due to Ahmet Hamdi Tanpınar being a writer who reviews his works a lot however this claims need to be backed up by Turkish literature experts.

Our observations on Ahmet Hamdi Tanpınar's works can be valuable for researchers working on Ahmet Hamdi Tanpınar and his works or historical linguistic and sociolinguistic analysis on Turkish language since interesting observations were done. They can also serve as evidence or support in literary criticism.

As a future study, the same methodology can be used to analyze other writer's works and other works of Ahmet Hamdi Tanpınar might be added analysis. Also, analysis of his essays and articles can be enlarged such that each essay and article whose date has the same year can be gathered to form text blocks with the label of the year. It would help us to see the real relationship between the average token length and the publication year for his essays and articles even though

the results are now helpful to support that our method provides correct results. Also, discriminant analysis might be added in terms of both type of the work and publication year of the work and classification can be done to some of the word blocks to see the success of the discriminant. In general, style analysis can be enlarged with new style measures and new methodologies.

Chapter 6

Conclusion and Future Work

In this thesis, we analyze Turkish texts from different aspects and drive conclusion about the methods to apply Turkish data. We use translation data, spoken texts or discourse, micro-blogs data and finally literary data, all in Turkish. Analyses applied use stylometric and machine learning methods.

In Chapter 2, the style similarity between original text of *My Name is Red* and its translations in English, French and Spanish is analyzed by using a rank consistency-based measure that we introduce in this thesis. The experiments with lexical features and most frequently used words statistically significantly confirm the similarity. They also show that the pairwise compatibility of translation languages is higher than that of their compatibility with the Turkish original. The observations are as expected for these target languages that are the members of the same language family. In future research, experiments with different language families can be performed. Furthermore, similar studies can be done with other works by dividing them into cohesive units, such as short stories or parts of novels containing related themes.

In Chapter 3, we provide a comprehensive quantitative analysis of spoken discourse for the Turkish language using memoirs of a group of old-time moviegoers. The aim is to infer various attributes of the participants from their speech.

Our experiments show that, to a large extent, age and gender groups can be attributed. We can accurately classify a text as written or spoken. Also, we see that the spoken language of a person changes itself to fit into the current spoken language of society, regardless of the age of the participant. Language use among age or gender groups differs in terms of the frequencies of the used words, rather than the words. Attributing participants to spoken texts from written texts is a challenging task. The classification accuracy in the experiments is not higher than a random guess. This shows that more information or mining is needed for the task. There are some future research possibilities that can be studied following this study. Additional language features can be investigated for analyzing groups. For example, n-gram usage differences can be further analyzed among age and gender groups. Other methods can be used for classification. Ensemble methods are more likely to give higher accuracy results, since they provide a classifier stronger than their components [71]. Movie choices of certain social, age, and gender groups can be investigated. Moreover, a social network can be created with the help of movie and movie theater names and the social networks of different groups can be compared. This may help us make sense of group choices and group dynamics of the past. Investigating the cultural heritage aspects of the memories is another path of research waiting to be done.

In Chapter 4, we use multiple Turkish micro-blog datasets for classification. We use English micro-blog datasets from the same and different domains, to investigate zero-shot cross-lingual transfer learning. We see that it is still not competitive in terms of its performance for transferring between distant languages. On the other hand, for large datasets coming from similar domains, cross-lingual experiments gives satisfactory results. Even though cross-lingual transferring from English to Turkish would not allow for an analysis without Turkish training data yet, it can help enriching the data at hand, or may help in data creation and annotation tasks.

In Chapter 5, we provide a stylometric analysis of Ahmet Hamdi Tanpınar's works. It is done by the calculation of six style measures: "sentence length in terms of words", "most frequent words", "word length of type", "word length of token", "syllable count of type" and "syllable count of token". PCA results

using these style measures show visual separation between the different types of works and between same types of works which has very different publication times. Those results lead us to make a type-wise analysis of the works. After doing so, we see that average token length relation with publication year changes from type to type. Our observations on Ahmet Hamdi Tanpınar's works can be valuable for researchers working on Ahmet Hamdi Tanpınar and his works or historical linguistic and sociolinguistic analysis on Turkish language since interesting observations are obtained. They can also serve as an evidence or support in literary criticism. As a future study, the same approach can be used to analyze other writers' works and other works of Ahmet Hamdi Tanpınar might be added. Also, analysis of his essays and articles can be enlarged such that each essay and article whose date has the same year can be gathered to form text blocks with the label of the year. Discriminant analysis might be added in terms of both type of the work and publication year of the work and classification can be done to some of the word blocks to see the success of the discriminant. In general, style analysis can be enlarged with new style measures and new methodologies.

Bibliography

[1] H. Akbulut, S. R. Öztürk, E. Uçar İlbuğa, and M. Gürer, "Kültürel ve toplumsal bir pratik olarak sinemaya gitmek: Türkiye'de seyirci deneyimleri üzerine bir sözlü tarih çalışması," tech. rep., Ankara, 2018.

[2] A. A. Akın, "Zemberek-NLP," 2019. Available at `https://github.com/ahmetaa/zemberek-nlp`, version 0.17.1.

[3] B. Diri and M. F. Amasyali, "Automatic author detection for Turkish texts," in *ICANN/ICONIP'03 13th International Conference on Artificial Neural Network and 10th International Conference on Neural Information Processing*, vol. 1, pp. 138–141, 2003.

[4] N. Saygılı, T. Amghar, B. Levrat, and T. Acarman, "Taking advantage of Turkish characteristic features to achieve authorship attribution problems for Turkish," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, 2017.

[5] P. Canbay, H. Sever, and E. A. Sezer, "Determining of discriminative blog size for authorship attribution on the Turkish texts," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, pp. 1–5, 2018.

[6] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing & Management*, vol. 44, no. 4, pp. 1448 – 1466, 2008.

[7] F. Can and J. M. Patton, "Change of word characteristics in 20th-century Turkish literature: A statistical analysis," *Journal of Quantitative Linguistics*, vol. 17, pp. 167–190, 2010.

[8] F. Can and J. M. Patton, "Change of writing style with time," *Computers and the Humanities*, vol. 38, pp. 61–82, 2004.

[9] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications," *ACM Comput. Surv.*, vol. 50, Nov. 2017.

[10] Wikipedia contributors, "Discourse analysis — Wikipedia, the free encyclopedia." `https://en.wikipedia.org/w/index.php?title=Discourse_analysis&oldid=940168511`, 2020. Online; accessed 11-February-2020.

[11] M. Corney, O. de Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *18th Annual Computer Security Applications Conference, 2002. Proceedings.*, pp. 282–289, Dec 2002.

[12] W. Dou, I. Cho, O. ElTayeby, J. Choo, X. Wang, and W. Ribarsky, "Demographicvis: Analyzing demographic information based on user generated content," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 57–64, Oct 2015.

[13] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data* (C. C. Aggarwal and C. Zhai, eds.), pp. 415–463, Boston, MA: Springer US, 2012.

[14] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," 2019.

[15] A. Eriguchi, M. Johnson, O. Firat, H. Kazawa, and W. Macherey, "Zero-shot cross-lingual classification using multilingual neural machine translation," *CoRR*, vol. abs/1809.04686, 2018.

[16] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-shot cross-lingual transfer with meta learning," in *EMNLP*, 2020.

[17] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, pp. 7059–7069, Curran Associates, Inc., 2019.

[18] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual sentiment analysis: An rnn-based framework for limited data," 2018. arXiv:1806.04511.

[19] A. Eriguchi, M. Johnson, O. Firat, H. Kazawa, and W. Macherey, "Zero-shot cross-lingual classification using multilingual neural machine translation," 2018. arXiv: 1809.04686.

[20] A. Mogadala and A. Rettinger, "Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 692–702, Association for Computational Linguistics, June 2016.

[21] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 451–462, Association for Computational Linguistics, July 2017.

[22] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 789–798, Association for Computational Linguistics, July 2018.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020. arXiv: 1911.02116.

[25] ICOMOS, *International Cultural Tourism Charter: Principles and Guidelines for Managing Tourism at Places of Cultural and Heritage Significance*. Burwood Victoria, Australia: ICOMOS International Cultural Tourism Committee, 2002.

[26] S. Çalışkan and F. Can, "Türkçe metinler üzerine yapılan sayısal üslup araştırmaları ve *Benim Adım Kırmızı* çevirilerinin aslına olan sadakatinin ölçümü," *Türk Kütüphaneciliği*, vol. 32, no. 4, pp. 251–286, 2018.

[27] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015.

[28] F. Can, E. F. Can, and C. Karbeyaz, "Translation relationship quantification: A cluster-based approach and its application to shakespeare's sonnets," in *Computer and Information Sciences* (E. Gelenbe, R. Lent, G. Sakellari, A. Sacan, H. Toroslu, and A. Yazici, eds.), (Dordrecht), pp. 117–120, Springer Netherlands, 2010.

[29] J. Patton and F. Can, "Determining translation invariant characteristics of James Joyce's *Dubliners*," in *Quantitative methods in corpus-based translation studies : a practical guide to descriptive translation research*, Amsterdam, Philadelphia: John Benjamins Pub., 2012.

[30] M. Baker, "Towards a methodology for investigating the style of a literary translator," *Target*, vol. 12, pp. 241–266, 05 2001.

[31] S. Çalışkan and F. Can, "Quantitative analysis of spoken discourse using memoirs of old-time moviegoers," *Journal of Quantitative Linguistics*, 2020 (Accepted for publication, December 22, 2020).

[32] K. Altintas, F. Can, and J. M. Patton, "Language change quantification using time-separated parallel translations," *Literary and Linguistic Computing*, vol. 22, pp. 375–393, 2007.

[33] G. L. Lewis, *The Turkish Language Reform: A Catastrophic Success*. Oxford University Press, 1999.

[34] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference.," *Journal of Personality and Social Psychology*, vol. 77 6, pp. 1296–312, 1999.

[35] C.-C. Li, R. M. Rodríguez, L. Martínez, Y. Dong, and F. Herrera, "Personalized individual semantics based on consistency in hesitant linguistic group decision making with comparative linguistic expressions," *Knowledge-Based Systems*, vol. 145, pp. 156 – 165, 2018.

[36] C. Thurlow and A. Brown, "Generation txt? The sociolinguistics of young people's text-messaging," *Discourse Analysis Online*, vol. 1, p. 30, 01 2003.

[37] H. Gómez-Adorno, D. Pinto, M. Montes, G. Sidorov, and R. Alfaro, "Content and style features for automatic detection of users' intentions in tweets," in *Advances in Artificial Intelligence – IBERAMIA 2014* (A. L. Bazzan and K. Pichara, eds.), pp. 120–128, Springer International Publishing, 2014.

[38] J. Karlgren and G. Eriksson, "Authors, genre, and linguistic convention," in *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, (Amsterdam, Netherlands), Association for Computing Machinery, 2007.

[39] D. Küçük and F. Can, "Stance detection: A survey," *ACM Comput. Surv.*, vol. 53, no. 1, p. 37, 2020.

[40] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, ""How old do you think I am?": A study of language and age in twitter," *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pp. 439–448, 01 2013.

[41] G. Redeker, "On differences between spoken and written language," *Discourse Processes*, vol. 7, no. 1, pp. 43–55, 1984.

[42] D. Biber, "Textual relations in speech and writing," in *Variation across Speech and Writing*, p. 121–169, Cambridge University Press, 1988.

[43] W. Chafe and D. Tannen, "The relation between written and spoken language," *Annual Review of Anthropology*, vol. 16, no. 1, pp. 383–407, 1987.

[44] N. Besnier, "The linguistic relationships of spoken and written nukulaelae registers," *Language*, vol. 64, no. 4, pp. 707–736, 1988.

[45] W. Horton, D. Spieler, and E. Shriberg, "A corpus analysis of patterns of age-related change in conversational speech," *Psychology and Aging*, vol. 25, pp. 708–713, 9 2010.

[46] G. Kavé, K. Samuel-Enoch, and S. Adiv, "The association between age and the frequency of nouns selected for production," *Psychology and Aging*, vol. 24 1, pp. 17–27, 2009.

[47] P. Eckert, *Age as a Sociolinguistic Variable*, ch. 9, pp. 151–167. John Wiley Sons, Ltd, 2017.

[48] R. Wodak and G. Benke, *Gender as a Sociolinguistic Variable: New Perspectives on Variation Studies*, ch. 8, pp. 127–150. John Wiley Sons, Ltd, 2017.

[49] M. A. Fitzpatrick, A. Mulac, and K. Dindia, "Gender-preferential language use in spouse and stranger interaction," *Journal of Language and Social Psychology*, vol. 14, no. 1-2, pp. 18–39, 1995.

[50] A. Hannah and T. Murachver, "Gender and conversational style as predictors of conversational behavior," *Journal of Language and Social Psychology*, vol. 18, no. 2, pp. 153–174, 1999.

[51] A. Hannah and T. Murachver, "Gender preferential responses to speech," *Journal of Language and Social Psychology*, vol. 26, no. 3, pp. 274–290, 2007.

[52] D. Cameron, *The myth of Mars and Venus: Do men and women really speak different languages?* Oxford University Press, 2008.

[53] J. Savoy, "Trump's and Clinton's style and rhetoric during the 2016 presidential election," *Journal of Quantitative Linguistics*, vol. 25, no. 2, pp. 168–189, 2018.

[54] D. Liu and L. Lei, "The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 us presidential election," *Discourse, Context Media*, vol. 25, pp. 143 – 152, 2018.

[55] Y. Wang and H. Liu, "Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election," *Discourse & Society*, vol. 29, no. 3, pp. 299–323, 2018.

[56] M. N. Coutanche and J. P. Paulus, "An empirical analysis of popular press claims regarding linguistic change in president Donald J. Trump," *Frontiers in Psychology*, vol. 9, p. 2311, 2018.

[57] C. Schoor, "In the theater of political style: Touches of populism, pluralism and elitism in speeches of politicians," *Discourse & Society*, vol. 28, no. 6, pp. 657–676, 2017.

[58] F. Pauli and A. Tuzzi, "The end of year addresses of the presidents of the Italian Republic (1948-2006): discoursal similarities and differences," *Glottometrics*, vol. 18, pp. 40–51, 2009.

[59] "World population prospects 2019." Online, `https://population.un.org/wpp/DataQuery/`, 2019.

[60] J. Rudman, "The state of authorship attribution studies: Some problems and solutions," *Computers and the Humanities*, vol. 31, no. 4, pp. 351–365, 1997.

[61] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H.

Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLOS ONE*, vol. 8, pp. 1–16, 09 2013.

[62] K. Oflazer and M. Saraçlar, *Turkish Natural Language Processing*, ch. 2 - Appendix: Turkish Morphological Features, pp. 46 – 51. Springer, 2018.

[63] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[64] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral Ecology*, vol. 26, no. 1, pp. 32–46.

[65] W. Haynes, *Holm's Method*, pp. 902–902. New York, NY: Springer New York, 2013.

[66] G. Drieman, "Differences between written and spoken language: An exploratory study," *Acta Psychologica*, vol. 20, pp. 36 – 57, 1962.

[67] M. Kestemont, E. Stamatatos, E. Manjavacas, W. Daelemans, M. Potthast, and B. Stein, "Overview of the cross-domain authorship attribution task at PAN 2019," in *CLEF*, 2019.

[68] H. Akbulut, "Bir seyirci araştırmasından etnografik deneyimler ve hikâyeler," *Folklor / Edebiyat*, vol. 24, pp. 13–34, 08 2018.

[69] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 1, pp. 9–26, 2009.

[70] C. Labbé and D. Labbé, "Inter-textual distance and authorship attribution Corneille and Molière," *Journal of Quantitative Linguistics*, vol. 8, no. 3, pp. 213–231, 2001.

[71] H. R. Bonab and F. Can, "GOOWE: Geometrically optimum and online-weighted ensemble classifier for evolving data streams," *ACM Trans. Knowl. Discov. Data*, vol. 12, Jan. 2018.

[72] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013. arXiv: 1309.4168.

[73] A. Lazaridou, G. Dinu, and M. Baroni, "Hubness and pollution: Delving into cross-space mapping for zero-shot learning," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Beijing, China), pp. 270–280, Association for Computational Linguistics, July 2015.

[74] I. Vulić and A. Korhonen, "On the role of seed lexicons in learning bilingual word embeddings," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 247–257, Association for Computational Linguistics, Aug. 2016.

[75] Y. Zhang, D. Gaddy, R. Barzilay, and T. Jaakkola, "Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1307–1317, Association for Computational Linguistics, June 2016.

[76] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," 2018. arXiv: 1710.04087.

[77] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov, "Emerging cross-lingual structure in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 6022–6034, Association for Computational Linguistics, July 2020.

[78] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," *CoRR*, vol. abs/1906.01502, 2019.

[79] M. F. Amasyali, H. Tasköprü, and K. Çaliskan, "Words, meanings, characters in sentiment analysis," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–6, 2018.

[80] A. Ucan, B. Naderalvojoud, E. A. Sezer, and H. Sever, "Sentiwordnet for new language: Automatic translation approach," in *2016 12th International*

Conference on Signal-Image Technology Internet-Based Systems (SITIS), pp. 308–315, 2016.

[81] Çöltekin, Çağrı, "A corpus of Turkish offensive language on social media," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 6174–6184, European Language Resources Association, May 2020.

[82] A. Bozyigit, S. Utku, and E. Nasibov, "Sanal zorbalık içeren sosyal medya mesajlarının tespiti," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 09 2018.

[83] E. Demirtas and M. Pechenizkiy, "Cross-lingual polarity detection with machine translation," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, (New York, NY, USA), Association for Computing Machinery, 2013.

[84] D. Küçük, "Stance detection in Turkish tweets," 2017. arXiv: 1706.06894.

[85] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), p. 649–657, MIT Press, 2015.

[86] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, (San Diego, California), pp. 31–41, Association for Computational Linguistics, June 2016.

[87] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 440–447, Association for Computational Linguistics, June 2007.

[88] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the ACL*, 2005.

[89] J. Liu, "515k hotel reviews data in europe." Online, `https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe`, 2017.

[90] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *Proceedings of NAACL*, 2019.

[91] S. Schweter, "BERTurk - BERT models for Turkish." Online: `https://doi.org/10.5281/zenodo.3770924` Version 1.0.0, Apr. 2020.

[92] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[93] A. Lüdeling and M. Kytö, *Corpus linguistics: An international handbook*. 03 2009.

[94] J. Patton and F. Can, "A stylometric analysis of Yaşar Kemal's "İnce Memed" tetralogy," *Computers and the Humanities*, vol. 38, pp. 457–467, 01 2004.

[95] "Hayatı." Online `http://www.tanpinarmerkezi.com/ahmet-hamdi-tanpinar/biyografi/`. Accessed: 2017-09-30.

[96] I. Enginün and Z. Kerman, *Günlüklerin Işığında Tanpınar'la Baş Başa*. Dergâh Yayınları, 2013.