Change of Word Characteristics in 20th-Century Turkish Literature: A Statistical Analysis

Fazli Can & Jon M. Patton

Routledge
Taylor & Francis Group

Change of Word Characteristics in 20th-Century Turkish Literature: A Statistical Analysis*

Fazli Can[1] and Jon M. Patton[2]

[1]Bilkent Information Retrieval Group, Computer Engineering Department, Bilkent University, Ankara, Turkey; [2]Information Technology Services, Miami University, Oxford, OH, USA

ABSTRACT

This article provides a century-wide quantitative analysis of the Turkish literature using 40 novels of 40 authors. We divide the century into four eras or quarter-centuries; allocate 10 novels to each era, and partition each novel into equal-sized blocks. Using cross-validation-based discriminant analysis, with the most frequent words as discriminators, we achieve a classification rate with a relatively high accuracy when the novel blocks are classified according to their eras. We show that, by using statistical stylistic methods, the author gender of Turkish texts can be accurately identified. We also study the gender differences regarding the use of most frequent words. Using weighted least squares regression and a sliding window approach we show that as time passes, words, both in terms of tokens (in text) and types (in vocabulary), have become longer. The findings of this work have implications for the historical linguistic analysis of the Turkish language.

1. INTRODUCTION

In quantitative natural language analysis, stylometric studies aim to discover hidden patterns or habits that are unconsciously used by authors (Stamatatos et al., 2000; Hoover, 2008). Such statistical stylistic methods are used in different fields that involve other kinds of human artefacts, such as architecture (Ozkar & Lefford, 2006), music (Backer & van Kranenburg, 2005), painting (Johnson et al., 2008), and programming (Ding & Samadzadeh, 2004). In stylistic text analyses, statistical methods employing

measurable text attributes (or style markers) – such as "most frequent words", "sentence lengths", and "word lengths" – are used (Forsyth & Holmes, 1996). The discovered hidden patterns, which conceptually correspond to fingerprints of authors, are used for various tasks such as authorship attribution (Smalheiser & Torvik, 2009), distinguishing works from each other according to intended audience (Binongo, 1994), genre detection (Kanaris & Stamatatos, 2009), author gender identification (Koppel et al., 2002), finding the chronological order of works (Stamou, 2007), or identifying an author's literary style development (Juola, 2007).

In this study we provide a quantitative analysis of the 20th-century Turkish literature using 40 novels of 40 different authors. We use an experimental approach aimed at quantifying patterns in the 20th-century Turkish literature and language in terms of various aspects of word usage. We investigate if author gender can be identified by their word usage patterns and quantify different aspects of the diachronic language change as time passes. To the best of our knowledge this is the only quantitative language analysis study that covers a full century.

In our analysis we divide the century into four eras or quarter-centuries, and include the full texts of 10 novels from each era. The reason for dividing the 100 years into four 25-year eras is that it provides a good framework for looking at temporal changes over time. It helps to eliminate variation between authors when we calculate the average of a style marker for each era. In our analysis we use six style markers: "most frequent words", "sentence length in terms of words", "syllable counts in tokens", "syllable count in types", "token length", and "type length". We identify the gender of the author of a given text using discriminant analysis. As suggested by Koppel et al. (2002) the problems we attack, i.e. era- and gender-based classifications, are more difficult than typical categorization and stylometry problems, since "individual authors are more likely to have consistent habits of style than are large classes of authors". We investigate the change in Turkish in terms of word lengths and the use of the most frequent words and (arguably) claim that language change may affect authors' word choices.

The findings of this work have implications for the historical linguistic analysis of the Turkish language. Its main contributions can be summarized as follows: in this paper we

- present the first quantitative investigation on Turkish covering a century based on a large corpus containing more than 2.5 million words;

- perform time- and gender-based text classifications in Turkish and show that they can be successfully achieved by using some commonly-used style markers;
- quantify temporal word length and frequent word usage change in Turkish in the 20th century;
- show that word length change may affect authors' common word choices.

2. RELATED WORK

In stylometry studies, writing styles of authors are analysed using objective measures. For this purpose about 1000 style markers have been identified (Rudman, 1998). The occurrence patterns of the selected style markers are used in various stylometric problems ranging from authorship attribution to gender detection. Similarly, the text categorization methods aim to assign texts into predefined categories such as known authors (Sebastiani, 2002).

A detailed overview of the stylometry studies in literature within a historical perspective is provided by Holmes (1994). It gives a critical review of numerous style markers and examines works on the statistical analysis of change of style with time. A critique of many authorship studies is provided by Rudman (1998). Stamou (2007) provides a survey of stylochronometric approaches of the last 60 years used for stylistic development analysis. Juola (2006) and Stamatatos (2009) provide a review of types of analysis, features, and recent developments in authorship attribution studies.

For analysing the occurrence patterns of style markers, various statistical methods are used. One popular technique is principal component analysis (PCA) (Binongo & Smith, 1999). Later in this work we use this statistical technique to visualize the separation between the works of the four eras. Another statistical technique we use in this study is discriminant analysis. Holmes and Forsyth (1995) use discriminant analysis to determine which vocabulary richness measures best discriminated between the Federalist papers written by Alexander Hamilton and those by James Madison. A similar work is undertaken by Bagavandas and Manimannan (2008) using canonical discriminant analysis to analyse three authors of Tamil language and show that their styles are clearly distinguishable from each other. Koppel et al. (2002)

exploit combinations of simple lexical and syntactic features to detect the gender of the author of a formal written text with about 80% accuracy. They use the same techniques to determine if a document is fiction or not with about 98% accuracy. Machine learning methods are also used in stylometric studies. For example, Koppel et al. (2009) study the cases that can occur in real-life authorship attribution applications. They define possible problems and show how machine learning techniques can be adapted to handle the special challenges of these problematic cases.

There are some Turkish-related recent works in the literature. For example, the writing style change of two Turkish authors, Yaşar Kemal and Çetin Altan, in their old and new works is analysed by Can and Patton (2004). In a more recent study, they analyse Yaşar Kemal's *İnce Memed* tetralogy using six style markers and provide accurate categorization results in discriminant analysis (Patton & Can, 2004). Kucukyilmaz et al. (2008) attack the categorization problem by using various machine learning algorithms for analysing Turkish chats in the Internet environment.

3. EXPERIMENTAL ENVIRONMENT

3.1 Turkish Language

In this study, by ''Turkish'' we mean the language mainly used in the republic of Turkey. Turkish belongs to the Altaic branch of the Ural-Altaic family of languages. Some concerns about this classification can be seen in Lewis (1988). The Turkish alphabet is based on Latin characters and has 29 letters consisting of 8 vowels and 21 consonants: *a, b, c, ç, d, e f, g, ğ, h, ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, z*. In some words borrowed from Arabic and Persian, the vowels ''*a*'', ''*i*'' and ''*u*'' are made longer or softer by using the character $^\wedge$ (circumflex accent) on top of them. In modern spelling, this approach is rarely used.

Turkish is an agglutinative language similar to Finnish and Hungarian. Such languages carry syntactic relations between words or concepts through discrete suffixes and have complex word structures. Turkish words are constructed using inflectional and derivation suffixes linked to a root. The meanings of the roots are enriched through affixation of derivational and inflectional suffixes. Like English, nouns in Turkish do not have a gender and the suffixes do not change depending on word type.

The study reported in Altintas et al. (2007) proposes and examines several stemming methods for Turkish by using a corpus that contains 712,272 tokens. The number of types (distinct words) in the corpus is 108,875, and the distinct number of stems for types is 24,388. The first 250 most frequent distinct stems constitute 47% of the corpus. Like other agglutinative languages, in Turkish it is possible to have words that would be translated into a complete sentence in non-agglutinative languages such as English. One well known example in this regard is the following Turkish word of 40 letters: *Avrupalılaştırılamıyabilenlerdenmişsiniz*, ''You seem to be one of those who may be incapable of being Europeanized'' (Lewis, 1988). However, the usage of such words is uncommon (Altintas et al., 2007; Can et al., 2008).

3.2 Corpus
In the construction of our corpus we aimed to cover the 20th-century Turkish literature in terms of its novels. We decided to divide the century into four quarters and include 10 novels from each era for a total of forty novels. Table 1 provides a listing of the novels, with some statistical information and the first publication year.

During corpus construction we aimed to represent both genders proportional to the real number of male and female authors in each era (Aksoy & Cankara, 2002). This implied a total of six female authors: one female author from each of the first two eras (namely Halide Edip Adıvar's *Kalb Ağrısı*, no. 7, from the first era; and Cahit Uçuk's *Dikenli Çit*, no. 13, from the second era) and two female authors from each of the last two eras (Adalet Ağaoğlu's *Ölmeye Yatmak*, no. 29; Füruzan's *Kırkyedililer*, no. 30, from the third era; and Pınar Kür's *Yarın Yarın*, no. 31; and Latife Tekin's *Sevgili Arsız Ölüm*, no. 35, from the fourth era). The authors and the works we covered are fairly representative in the history of the 20th-century Turkish novel (Naci, 1999; Necatigil, 1992).

The novel sizes, in terms of number of words, vary between 25,201 (Oktay Akbal's *Garipler Sokağı*, no. 20) and 159,724 (Oğuz Atay's *Tutunamayanlar*, no. 28) and have an average of 63,499 and a standard deviation of 31,474 words. As shown in Table 2, the total text sizes of era 1 to era 4, in order, are 542,645; 520,924; 796,303; and 680,125 words, and all together 2,539,997 words. The publication information for the novels used for optical character recognition during digitization is provided in the Appendix.

Table 1. Authors and their novels with some statistical information (+: female author's work).*

No.	Author, novel, first publication year	No. of tokens (N)	No. of types (V)	Average token length	Average type length	Average sentence length
1	Mehmet Rauf, *Eylül*, 1901	68,899	14,825	5.906	8.264	14.550
2	Hüseyin Rahmi Gürpınar, *Toraman*, 1919	34,479	12,904	6.088	7.854	10.550
3	Ömer Seyfettin, *Efruz Bey*, 1919	41,828	14,362	6.245	8.208	6.778
4	Refik Halit Karay, *İstanbul'un Bir Yüzü*, 1920	39,275	14,258	6.071	7.950	13.460
5	Reşat Nuri Güntekin, *Çalıkuşu*, 1922	90,610	20,842	6.121	8.375	9.384
6	Yakup Kadri Karaosmanoğlu, *Nur Baba*, 1922	34,608	10,424	5.774	7.763	13.570
7+	Halide Edip Adıvar, *Kalb Ağrısı*, 1924	55,981	14,609	5.995	8.032	11.350
8	Salâhattin Enis, *Zaniyeler*, 1924	47,303	13,973	6.039	8.116	10.590
9	Halit Ziya Uşaklıgil, *Kırık Hayatlar*, 1924	83,451	19,472	6.115	8.401	15.480
10	Peyami Safa, *Sözde Kızlar*, 1925	46,211	13,116	6.122	8.157	8.540
11	Mahmut Yesari, *Tipi Dindi!*, 1933	45,876	13,818	6.156	8.053	6.813
12	M. Şevket Esendal, *Ayaşlı İle Kiracıları*, 1934	47,552	11,425	5.710	7.985	7.252
13+	Cahit Uçuk, *Dikenli Çit*, 1937	33,754	9902	6.273	8.098	6.563
14	Mithat Cemal Kuntay, *Üç İstanbul*, 1938	129,568	27,291	6.364	8.525	10.150
15	Abdülhak Şinasi Hisar, *Fahim Bey ve Biz*, 1941	34,513	11,613	6.138	8.290	16.340
16	Sabahattin Ali, *Kürk Mantolu Madonna*, 1943	39,969	11,548	6.213	8.406	9.261
17	Kemal Bilbaşar, *Denizin Çağırışı*, 1943	29,755	11,508	6.406	8.316	11.550
18	Sait Faik, *Medarı Maişet Motoru*, 1944	36,075	12,011	6.037	7.897	7.741
19	Ahmet Hamdi Tanpınar, *Huzur*, 1949	98,661	23,419	6.159	8.545	9.691
20	Oktay Akbal, *Garipler Sokağı*, 1950	25,201	8849	6.310	8.062	9.085
21	Orhan Kemal, *Cemile*, 1952	28,256	9364	5.894	7.579	5.714

Table 1. (*Continued*).

No.	Author, novel, first publication year	No. of tokens (N)	No. of types (V)	Average token length	Average type length	Average sentence length
22	Yaşar Kemal, *İnce Memed* [1], 1955	86,996	17,113	5.772	8.015	5.348
23	Yusuf Atılgan, *Aylak Adam*, 1959	35,930	11,451	6.150	8.057	5.279
24	Kemal Tahir, *Yorgun Savaşçı*, 1965	100,331	26,592	6.310	8.591	5.834
25	Tarık Buğra, *İbiş' in Rüyası*, 1970	48,931	13,853	5.913	8.192	7.820
26	Fakir Baykurt, *Tırpan*, 1970	88,887	19,742	5.716	7.757	4.567
27	Çetin Altan, *Büyük Gözaltı*, 1972	44,023	12,685	6.148	8.149	7.218
28	Oğuz Atay, *Tutunamayanlar*, 1972	159,724	37,676	6.404	8.964	7.653
29+	Adalet Ağaoğlu, *Ölmeye Yatmak*, 1973	86,353	25,332	6.118	8.442	6.973
30+	Füruzan, *Kırkyedililer*, 1974	116,872	35,103	6.642	8.981	7.560
31+	Pınar Kür, *Yarın Yarın*, 1976	83,861	21,178	6.147	8.705	8.420
32	Ferid Edgü, *O; Hâkkari'de Bir Mevsim*, 1977	26,145	8,640	5.934	7.885	5.860
33	Selim İleri, *Ölüm İlişkileri*, 1979	52,812	17,225	6.621	8.801	9.902
34	Orhan Pamuk, *Sessiz Ev*, 1983	79,236	18,133	5.957	8.417	9.099
35+	Latife Tekin, Sevgili *Arsız Ölüm*, 1983	59,010	12,932	6.169	7.906	7.958
36	Mehmet Eroğlu, *Issızlığın Ortasında*, 1984	71,673	17,794	6.390	8.538	6.280
37	Attilâ İlhan, *Haco Hanım Vay!*, 1984	107,671	31,320	6.423	8.434	11.05
38	Kaan Arslanoğlu, *Devrimciler*, 1987	74,376	19,913	6.199	8.589	6.544
39	Nedim Gürsel, *Boğazkesen:Fatih'in Romanı*, 1995	52,393	18,090	6.531	8.435	10.19
40	Ahmet Altan, *Kılıç Yarası Gibi*, 1998	72,948	20,649	6.447	8.668	20.33

*Researchers who are interested in doing research with the same dataset can request a copy of the corpus from the authors of this paper.

In this study an individual text word, *token*, is defined as a continuous string of word characters. The "word" characters are the Turkish letters, additional letters from the English alphabet (*q, x, w*), dash ("-" used in
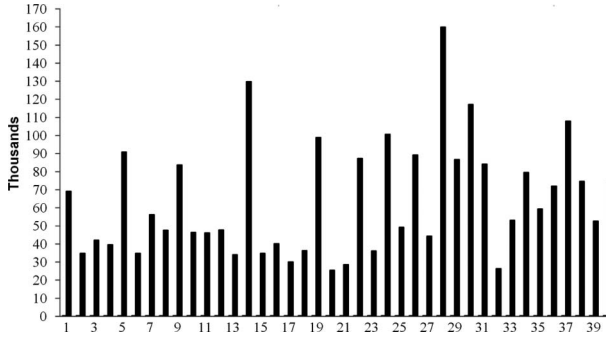
Fig. 1. Novel sizes (*y*-axis values: in terms of number of words) in graphical form, min: 25,201 (no. 20); max: 159,724 (no. 28); average: 63,499; standard deviation: 31,474 words.

Table 2. Number of tokens, types, and their length information for eras.

Era: novels	No. of tokens (N)	No. of types (V)	Average token length	Average type length
Era 1: novels 1–10	542,645	80,739	6.055	8.849
Era 2: novels 11–20	520,924	76,411	6.192	9.018
Era 3: novels 21–30	796,303	113,834	6.176	9.246
Era 4: novels 31–40	680,125	100,327	6.292	9.262
All novels 1–40	2,539,997	233,118	6.185	9.575

old words), apostrophe sign, and the digits from 0 to 9. The versions of the letters "*a*", "*i*", and "*u*" with a ^ (circumflex) on top of them are included and regarded as different from "*a*", "*i*" and "*u*". A *type* is defined as a distinct word. The term *vocabulary* (or *lexicon*) means the set of all types. According to our definition a word begins with a letter and ends with a non-word character, and the case of letters is insignificant. The minimum word length is defined as two (word) characters; however, the only one-letter word of Turkish "*o*" (means "he, she, it") was counted as a word.

4. EXPERIMENTAL DESIGN

4.1 Selection of Block Size and Style Markers
For most of our discriminant analyses and ANOVAs (analysis of variance) we needed observations based on fixed-size text blocks. We decided

that 2500 is an appropriate block size to be used (Can & Patton, 2004). In this way we were also able to obtain at least 10 blocks from each novel. We only use complete blocks and the incomplete residual last blocks of the novels are not used in the experiments. Thus, for the eras 1 to 4, we respectively obtained 211, 203, 312, and 266 blocks (all together 992 blocks).

As indicated earlier we used six style markers: "most frequent words", "sentence length in terms of words", "syllable counts in tokens", "syllable count in types", "token length" and "type length". For sentence length we counted the number of words in each sentence. Any sentence with a length of 250 and more is assumed to have the length 250. There were very few sentences like this. However, the novel *Tutunamayanlar* (no. 28) had eight sentences with more than 1000 words (actually in these passages, as a part of his literary style, Oğuz Atay writes without end of sentence punctuation symbols). Orhan Pamuk's novel *Sessiz Ev* (no. 34) contains two sentences with more than 1000 words. Like that of Atay's case, this is a part of Pamuk's literary style. Perhaps by this way these authors are trying to present their capabilities in literary art. Any word with a length of 30 and more is assumed to have the length 30. Such words are uncommon in Turkish and appear only a few times (Can et al., 2008). Again Oğuz Atay's novel contains words like this, since in this novel some sentences are in the form of single words without the use of blank spaces.

4.2 Determining Most Frequent Words

For the selection of most frequent words we wanted to give the same significance to each novel independent of its size. For this purpose we found the relative frequency of each word (i.e. the ratio of number of occurrences of a word to the total number of tokens) in each novel, summed these relative frequencies, sorted the words according to this sum, and selected the top 30 (in some experiments, the top 60) words. In the experiments we use these words as our most frequent words. These words are listed in Figure 2 in two different ways: first according to decreasing order of occurrence frequencies and then for easy reference in alphabetical order.

5. EXPERIMENTAL RESULTS

In this section we first perform a principal component analysis (PCA) and cluster the novels using most frequent words to graphically illustrate the differences among the works of four eras. PCA is used as a motivational

bir (a, an, one), *bu* (this), *ve* (and), *de* (too), *o* (he, she), *ne* (what), *da* (too), *gibi* (like), *sonra* (later), *için* (for), *kadar* (until), *daha* (more), *ben* (I), *dedi* (said), *diye* (that), *her* (every), *ki* (that, which, who), *çok* (very), *ama* (but), *hiç* (never, none, nothing), *mi* (adverb of interrogation), *iki* (two), *değil* (not), *fakat* (but), *bütün* (all), *onun* (hers, his, its), *onu* (her, him, it), *var* (there is, there are), *şey* (thing), *zaman* (time), *beni* (me), *bana* (to me), *içinde* (in), *mı* (adverb of interrogation), *böyle* (so), *bile* (even), *ile* (and), *gün* (day), *şimdi* (now), *nasıl* (how), *sen* (you), *ya* (then, so), *yok* (there is no), *en* (most), *bey* (gentleman, Mr.), *benim* (my), *ona* (to her, to him), *biraz* (a little), *başka* (other), *artık* (any more), *vardı* (there was, there were), *kendi* (herself, himself), *öyle* (so), *büyük* (large), *iyi* (good), *belki* (perhaps), *yalnız* (only), *küçük* (small), *kadın* (woman), *doğru* (right)

A

ama (but), *artık* (any more), *bana* (to me), *başka* (other), *belki* (perhaps), *ben* (I), *beni* (me), *benim* (my), *bey* (gentleman, Mr.), *bile* (even), *bir* (a, an, one), *biraz* (a little), *bu* (this), *böyle* (so), *bütün* (all), *büyük* (large), *çok* (very), *da* (too), *daha* (more), *de* (too), *dedi* (said), *değil* (not), *diye* (that), *doğru* (right), *en* (most), *fakat* (but), *gibi* (like), *gün* (day), *her* (every), *hiç* (never, none, nothing), *için* (for), *içinde* (in), *iki* (two), *ile* (and), *iyi* (good), *kadar* (until), *kadın* (woman), *kendi* (herself, himself), *ki* (that, which, who), *küçük* (small), *mı* (adverb of interrogation), *mi* (adverb of interrogation), *nasıl* (how), *ne* (what), *o* (he, she), *ona* (to her, to him), *onu* (her, him, it), *onun* (hers, his, its), *öyle* (so), *sen* (you), *sonra* (later), *şey* (thing), *şimdi* (now), *var* (there is, there are), *vardı* (there was, there were), *ve* (and), *zaman* (time), *ya* (then, so), *yalnız* (only), *yok* (there is no)

B

Fig. 2. Most frequent 60 words (each word is followed by its English translation). (A) Words are listed in decreasing order of occurrence frequencies. (B) Words are listed in alphabetical order; "*da*" and "*de*" are essentially the same word (the surface form difference is due to vowel harmony).

tool for additional analyses as well as a visualization tool. Next, for each style marker, a stepwise discriminant analysis is conducted to determine the best discriminators for classifying novel blocks to their respective eras and in effect reduces the dimensionality of the problem. This is followed by a discriminant analysis using cross-validation to determine the success rate of these discriminators. In cross-validation, each block of a given novel is classified based on linear discriminant functions derived from blocks of all the other novels. These linear discrimination functions are derived using the best discriminators obtained from a stepwise discriminant analysis applied to all the blocks of the other novels. This strict form of cross-validation completely eliminates author bias.

We also consider the author gender identification problem using discriminant analysis and studying the gender differences regarding the use of most frequent words. Then we study the change in the usage of most frequent words independent of gender. Finally, using blocks, we compare the average token and type lengths among the works of each era by conducting an ANOVA.

5.1 Clustering of Novels Using Style Markers

The PCA plot, Figure 3, provides a visualization of the similarities among the novels by using the number of occurrences information of the
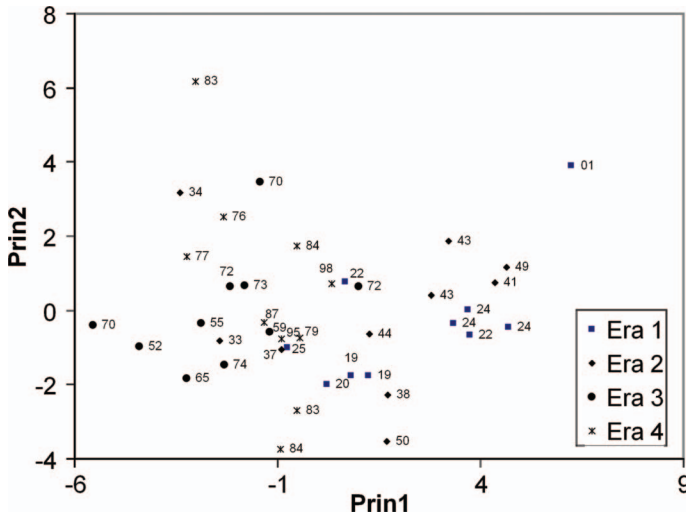
Fig. 3. PCA results using top 30 words and one vector for each novel.

most frequent 30 words. Using the top 30 was enough since the first two components explain 39% of the variance of the terms and using 60 does not make it any higher. For PCA we use the complete novels' vectors (i.e. no blocking). In this plot the novel symbols represent one of the four eras in which the novel was written. The labels represent the difference between the publication year and 1900. Note that novels from the first and second era are mostly placed on the right-hand side and the novels of the other two eras are located on the left. Also note that the three novels written in 1924 (denoted by the data label 24) and one of the novels written in 1922 (Yakup Kadri Karaosmanoğlu's *Nur Baba*, the novel no. 6, in the plot denoted by the label 22) have data points very close to each other. The same is true for the two novels written in 1919 and the one written in 1920. Also the two written in 1943 are very close.

The novels of the third era (novel no./publication year: 21/1952, 22/1955, 23/1959, 24/1965, 26/1970, 27/1972, 29/1973, 30/1974) are again placed very close to each other. Mehmet Rauf's *Eylül* (no. 1, published in 1901) and Orhan Pamuk's *Sessiz Ev* (no. 34, published in 1983) are located far from the other works.

The PCA applied to the other style markers did not provide the separation among the eras as the most frequent words.

5.2. Classifying Novels According to Eras with Style Markers
The details of classification for all style markers are given in Table 3. Discriminant analysis using sentence lengths has an (average) success rate of 34.81% for four eras. This is better than 25% success rate if everything is random. When the token lengths are used, the success rate is 32.16%. For type lengths the success rate is 37.71%. Syllable counts in types have a success rate of 37.52%. Syllable counts in tokens have a success rate of 34.67%. All of these results are obtained using blocks. These markers are not good enough for significantly clear discrimination since they show change even among the authors of the same era. With more novels one may obtain better results.

On the other hand, discriminant analysis using most frequent words gave results with high accuracy in distinguishing the four eras. Using 60 most frequent words, a success rate of 57.27% was achieved using cross-validation.

5.3 Gender Identification and Frequent Word Usage Change with Gender
In this section we consider two gender-related problems. In the analysis of these problems it should be remembered that, unlike some other Asiatic languages such as Arabic and Japanese, Turkish is gender-neutral, i.e., involves no gender-specific constructs.

5.3.1 Gender Identification
Using the 40 complete novel vectors, a series of discriminant analyses were conducted for classifying gender using various style markers. To eliminate author bias, cross-validation was used where each novel was classified based on linear discrimination functions derived from the other novels.

Using the 30 most frequent words provides a correct weighted classification rate of 94.1%. In this case, all female authors (6) are classified correctly, whereas four male authors are classified incorrectly. In this calculation, male and female authors are weighted equally when we use "weighted classification". In our case the 100% correct classification rate of female authors is averaged with the $30/34 = 88.23\%$ classification rate of male authors (the average of these two numbers provide the correct weighted classification rate given above). The average word frequency for each novel was used as the data values where the average was based on word frequencies of each block of a given novel. The best discriminators for the 30 most frequent word cases are the following: "*bu*", "*çok*", "*da*", "*ne*", "*ama*", "*için*", and "*kadar*".

Table 3. Correct (diagonal, in bold) and incorrect (off-diagonal) classification rates for each style marker using blocks. Average correct classification rate for most frequent 60 words: 57.27%; sentence length: 34.10%; syllable count in tokens: 34.67%; syllable count in types: 37.52%; token lengths: 32.16.%; type lengths: 37.71%.

Era (no. of blocks)	Style marker	E1		E2		E3		E4	
E1 (211)	Most frequent 60 words	**72.00%**	(152)	27.01%	(57)	0.95%	(2)	0.00%	(0)
	Sentence lengths	**46.92%**	(99)	29.38%	(62)	4.26%	(9)	19.43%	(41)
	Syllable count in tokens	**44.54%**	(94)	32.70%	(69)	17.54%	(37)	5.21%	(11)
	Syllable count in types	**44.08%**	(93)	32.70%	(69)	15.17%	(32)	8.06%	(17)
	Token lengths	**32.23%**	(68)	29.86%	(63)	19.91%	(42)	18.01%	(38)
	Type lengths	**45.02%**	(95)	37.44%	(79)	12.32%	(26)	5.21%	(11)
E2 (203)	Most frequent 60 words	33.99%	(69)	**45.32%**	(92)	18.72%	(38)	1.97%	(4)
	Sentence lengths	30.05%	(61)	**24.14%**	(49)	22.17%	(45)	23.65%	(48)
	Syllable count in tokens	25.62%	(52)	**39.90%**	(81)	18.23%	(37)	16.26%	(33)
	Syllable count in types	30.54%	(62)	**47.78%**	(97)	10.34%	(21)	11.33 %	(23)
	Token lengths	34.00%	(69)	**34.98%**	(71)	15.27%	(31)	15.76%	(32)
	Type lengths	33.50%	(68)	**42.86%**	(87)	12.32%	(25)	11.33%	(23)
E3 (312)	Most frequent 60 words	0.00%	(0)	12.18%	(38)	**53.85%**	(168)	33.97%	(106)
	Sentence lengths	8.01%	(25)	27.56%	(86)	**52.56%**	(164)	11.86%	(37)
	Syllable count in tokens	27.24%	(85)	16.99%	(53)	**21.15%**	(66)	34.62%	(108)
	Syllable count in types	22.12%	(69)	16.67%	(52)	**28.53%**	(89)	32.69%	(102)
	Token lengths	22.76%	(71)	20.19%	(63)	**29.49%**	(92)	27.56%	(86)
	Type lengths	16.99%	(53)	13.46%	(42)	**38.14%**	(119)	31.41%	(98)
E4 (266)	Most frequent 60 words	1.13%	(3)	0.75%	(2)	40.23%	(107)	**57.89%**	(154)
	Sentence lengths	28.95%	(77)	28.20%	(75)	30.08%	(80)	**12.78%**	(34)
	Syllable count in tokens	15.79%	(42)	21.05%	(56)	30.08%	(80)	**33.08%**	(88)
	Syllable count in types	13.91%	(37)	20.30%	(54)	36.09%	(96)	**29.70%**	(79)
	Token lengths	31.20%	(83)	9.40%	(25)	27.44%	(73)	**31.95%**	(85)
	Type lengths	16.92%	(45)	21.43%	(57)	36.84%	(98)	**24.81%**	(66)

Another style marker providing highly accurate results in discriminating gender was sentence length. The weighted classification rate using cross-validation was 94.1% with all female authors classified correctly and only four male authors classified incorrectly. Sentences of eight, nine or 10 words were the best discriminators, with females writing a higher percentage of these. A sentence of 39 words was also a good discriminator, with male authors writing a higher percentage.

Other style markers providing good results were token lengths providing a weighted classification rate of 83.8% and type lengths providing a rate of 77.5%.

The discriminant analysis using the other style markers, syllable counts in tokens and syllable counts in types, were less impressive providing classification rates of 52% and 57% respectively.

5.3.2 Frequent Word Usage Change with Gender

Using blocks as the unit of measurement, a MANOVA was conducted to determine which of the most frequent words had a significant change in usage during the four eras for each gender. The frequency of the best word discriminators for gender was used as the response variables, and the era number was used as the classification variable. Separate analyses were done for each gender. Multiple comparisons of the frequencies of these words between eras were done using a Scheffe correction since six such comparisons were required to find usage change. Table 4 lists the

Table 4. Most frequent words whose usage significantly changed from era to era for each gender.

Era comparison	Gender	Most frequent words
E1–E2	M	ama, bu, çok, da
	F	bu, için, kadar
E1–E3	M	ama, bu, çok, da, için, kadar
	F	ama, bu, da, için, kadar
E1–E4	M	ama, bu, da, çok, kadar
	F	ama, bu, da, için, kadar
E2–E3	M	ama, bu, da, için, kadar
	F	ama, da, kadar
E2–E4	M	ama, bu, da, için, kadar
	F	ama, da, için
E3–E4	M	ama, bu, da, kadar
	F	ama, bu, da, için, kadar

most frequent words whose usage has significantly changed between eras for both male and female authors.

For male authors the usage of "*çok*", "*da*" and "*ama*" has increased from Era 1 to Era 4, whereas the usage of the word "*bu*", "*için*" and "*kadar*" have decreased. For female authors the words "*da*" and "*ama*" have increased usage over the four eras whereas "*bu*", "*çok*", "*için*" and "*kadar*" have decreased.

5.4 Identifying Language Change in Terms of Frequent Word Use and Word Length

5.4.1 Change in Most Frequent Words and Word-Choice Preference Change

The use of words may also show change with time in terms of their usage frequency. For example, Woods (2003) shows that the most frequent word in modern Spanish was considerably less frequent during the 16th and 17th centuries. Table 5 lists words whose frequency of usage is significantly different between eras independent of gender. The words showing increasing usage in between Era 1 and Era 4 are "*de*", "*ama*" and "*da*"; the words showing decreasing usage are "*bir*", "*bu*", "*bütün*", "*fakat*," "*kadar*" and "*onun*".

We do get strong relationships with frequency usage of "*ama*" and "*fakat*" versus both year and era using blocks as the unit of measurement. The usage of "*ama*" increases and "*fakat*" decreases over time. For example, in terms of blocks the correlation is –0.1022; it is significant at $p = 0.0013$. These words mean "but" in English. We claim that less frequent use of "*fakat*" and more frequent use of "*ama*" or the replacement of the former with the latter is due to increasing word lengths with time (see the next section for the discussion of the temporal word length increase in Turkish). Since Turkish words have become longer with time, to add a variety to their style as time passes, authors are

Table 5. Most frequent words whose usage significantly changed from era to era.

Era comparison	Most frequent words
E1–E3	ama, bir, bu, da, de, fakat, kadar
E1–E4	ama, bu, bütün, da, fakat, kadar, onun
E2–E3	bu, de, fakat
E2–E4	ama, bu, fakat, onun

inclined to prefer "*ama*" (a shorter word) over "*fakat*" (a longer word) with the same meaning. Note that both of these words are loan words that are borrowed from Arabic (TDK, 1974). Therefore, this preference shift cannot be attributed to the Turkish language (purification) reform, which aimed to replace words of foreign origin with their Turkish equivalents (Lewis, 1999). Further research is needed on this; however, the strong statistical relationship we observe is an important pointer on the validity of this claim.

5.4.2 Change in Word Lengths

To illustrate comparisons of values for the style markers, type length and token length for each era, two plots are presented in Figure 4. The first (left) scattergram based on 40 novel vectors (no blocking) shows that token length tends to increase as the publication date increases. A weighted least squares regression was performed. Since the sizes of the novels vary from one another, the weight applied to each novel was based on its total number of tokens. For this regression $R^2 = 0.153$, $F(1,38) = 6.87$ ($p = 0.012$). Since $p < 0.05$, we have significant evidence of a linear relationship. This is truly remarkable considering the large amount of intra-author variation between novels and our modest sample size. The prediction equation is given by

$$\text{Average token length} = 5.9741 + 0.00368 * (\text{year} - 1900)$$

So for each year after 1900, the average token length would increase by approximately 0.00368. The predicted values are included in the plot.
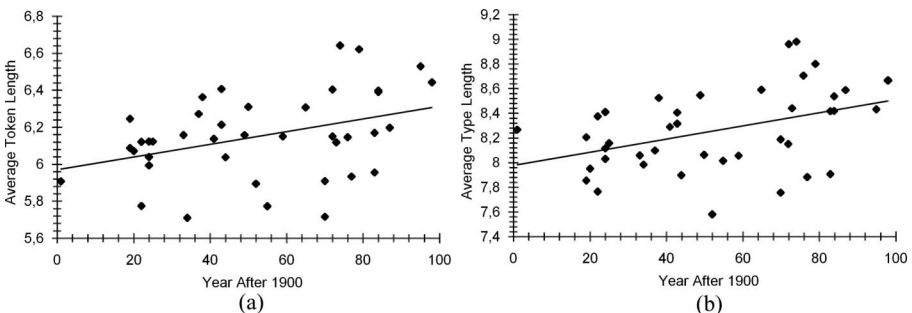

Fig. 4. Scattergram of average token and type length versus the year of publication. (a) Scattergram of average token length versus the year of publication. (b) Scattergram of average type length versus the year of publication.

Likewise, type lengths tend to increase over time according to the second (right) scattergram of Figure 4 based on 40 novel vectors. Again, we applied weighted least squares regression using the same weights as the above analysis. For this regression $R^2 = 0.200$, $F(1,38) = 9.50$ ($p = 0.004$). Since $p < 0.01$, the results indicate strong evidence of a linear relationship. The prediction equation is given by

$$\text{Average type length} = 8.0127 + 0.00609 * (\text{year} - 1900)$$

So for each year after 1900, the average type length would increase by approximately 0.00609.

We also use a sliding window approach that exploits smaller-scale temporal groupings (Kjell et al., 1994). Its purpose is to capture token and type length variations with a finer level of granularity (Abbasi & Chen, 2008). In this approach five consecutive novels were used for each data point. The first data point was based on the first five novels. The second point was based on novels 2 to 6 and so on; novels 36 to 40 formed the last data point. This gave us a total of 36 observations. Similar results were obtained for both types and tokens using the ''sliding window'' approach to the regression. The token length was the weighted average of the respective average token lengths of the five novels where the weights were based on the respective number of tokens of the five novels. Likewise, the type length was a weighted average using the number of types of each novel as the basis for the weights. The year coordinate was the average of the publication years of the five novels. The regression equations for token and type lengths were almost identical to those above. However, the results were much more significant since much of the variation between novels was smoothed out due to the averaging.

Even more powerful results were obtained using blocks as cases instead of the complete novel vectors. ANOVAs were conducted using a nested design where the classification variable, novel number (1–40), was nested in the classification variable era. Token length and type length were the response variables. The token length ANOVA had an $R^2 = 0.767$, an F statistic $F(39,954) = 80.32$ ($p < 0.0001$), and the average token lengths for each era is shown in Table 6. With the exception of eras 2 and 3, each pairwise comparison was significant at an $\alpha = 0.05$.

The type length ANOVA had an $R^2 = 0.763$, an F statistic $F(39,954) = 78.56$ ($p < 0.0001$), and average type lengths for each era

Table 6. Average token and type length for each era using blocks.

Era	Average token length	Average type length
1	6.055	7.177
2	6.192	7.272
3	6.176	7.227
4	6.295	7.352

are shown in Table 6. Note that the average type length for each era is the average of all average type lengths for each block belonging to that era. With the exception of eras 1 and 3, each pairwise comparison was significant at ($\alpha < 0.05$).

Both the average token and type length in the 3rd era were smaller than the respective lengths in eras 2 and 4 due to the inclusion of three village novels (21, 22, and 26) in era 3. The fictional characters of these novels are mostly rural people and the novels contain dialogs among them.

One can argue that for measuring word length change we can use sample texts from each quarter-century, compute the mean word length, test for normality, and then carry out a significance test on the results. To perform such an analysis, we should use the entire set of 40 novels to determine how type or token lengths vary over time. By choosing one novel at random from each era, and using blocks as a unit of measurement, one would be able to analyse how token lengths vary but not type lengths (since in that case the entire novel would need to be used as a measurement unit). Using all 40 novels would provide a richer set of data. By choosing 10 novels from each era or using the sliding time window approach we also get a balanced design.

6. CONCLUSIONS AND FUTURE WORK

We provide a quantitative analysis of the 20th-century Turkish literature using a stylometric approach. The PCA results using most frequent words provide visual separation between the works of different eras. These results motivate further analysis. We use discriminant analysis based on cross-validation using six style markers. Among these style markers, "most frequent words" provide the best discrimination among the four eras. Sentence lengths, as well as most frequent words, provide

decidedly accurate (more than 90%) separation of authors according to their gender. As also indicated by Koppel et al. (2002) this is a surprising result, since large classes of authors are likely to exhibit inconsistent habits of style.

Our findings show that in Turkish as time passes, words, both in terms of tokens and types, have become longer. These are in agreement with the findings of our previous studies which are based on parallel old and new translations of seven works (Altintas et al., 2008) and old and new works of two Turkish authors (Can & Patton, 2004). The results presented in this work are much more widespread and decisive, since they are based on a much larger corpus that includes the works of, not two, but several different authors. They also cover a century-wide time span and involve no translations. In measuring the word length change, we not only use the four eras but also a sliding window approach. This approach captures token and type length variations with a finer level of granularity since five consecutive novels are used to generate each data point. In the sliding window approach, the results are much more significant since many of the variations between novels are smoothed out due to the averaging.

The increase in word lengths with time can be attributed to the government-initiated language "reform" of the 20th century (Lewis, 1999). This reform aimed at replacing foreign words used in Turkish, especially Arabic- and Persian-based words (since they were in the majority when the reform was initiated in early 1930s), with newly-coined pure Turkish neologisms created by adding suffixes to Turkish word stems (Lewis, 1999).

Based on our observations of the change of a specific word use (more specifically in newer works the preference of "*ama*" over "*fakat*" where both mean "but" in English and their inverse usage correlation is statistically significant), we speculate that the word length increase can influence the common word choice preferences of authors. It should be noted that these words ("*ama*" and "*fakat*") are borrowed from Arabic and, therefore, the preference change cannot be attributed to the Turkish language reform.

Our observations on Turkish language change are valuable for researchers working on historical linguistic or sociolinguistic analysis of the Turkish language. The principal component analysis-based clustering results similar to ours can be used for identifying previously unknown similarities and differences of the eras and authors' styles. They can also serve as external evidence in literary criticism. Our gender-related

findings can be useful in gender-specific studies in Turkish literature. In future studies, machine learning techniques or syntax- and character-based measures can be used. Furthermore, temporal groupings of novels can be done using taxonomies from literary studies or important time marks. The effects of word length change on authors' common word choices and further quantitative analysis of temporal word choice preference shifts are other interesting future research possibilities.

ACKNOWLEDGEMENTS

REFERENCES

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, *26*(2).

Aksoy, E., & Cankara, M. (2002). Çağdaş Türk edebiyatçısının toplumsal profili. *Kanat Bilkent Üniversitesi Türk Edebiyatı Merkezi Haber Bülteni*, No. 10. Retrieved November 15, 2009, from http://www.bilkent.edu.tr/∼kanat/profil.html

Altintas, K., Can, F., & Patton, J. M. (2007). Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing*, *22*(4), 375–393.

Backer, E., & van Kranenburg, P. (2005). On musical stylometry: A pattern recognition approach. *Pattern Recognition Letters*, *26*(3), 299–309.

Bagavandas, M., & Manimannan, G. (2008). Style consistency and authorship attribution: A statistical investigation. *Journal of Quantitative Linguistics*, *15*(1), 100–110.

Binongo, J. N. G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: A statistical expression. *Literary and Linguistic Computing*, *9*(4), 267–279.

Binongo, J. N. G., & Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, *14*(4), 445–465.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., & Vursavas, O. M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, *59*(3), 407–421.

Can, F., & Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities*, *38*(1), 61–82.

Ding, H., & Samadzadeh, M. H. (2004). Extraction of Java program fingerprints for software authorship identification. *Journal of Systems and Software*, *72*(1), 49–57.

Forsyth, R. S., & Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, *11*(4), 163–174.

Holmes, D. I., & Forsyth, R. S. (1995). The *Federalist* revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, *10*(2), 111–127.

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

Hoover, D. L. (2008). Quantitative analysis and literary studies. In R. Siemens & S. Schreibmann (Eds), *A Companion to Digital Literary Studies* (pp. 517–533). Oxford: Blackwell.

Johnson, C. R., Hendriks, E., Berezhnoy, I., Brevdo, E., Hughes, S. M., Daubechies, I., Li, J., Postma, E., & Wang, J. Z. (2008). Image processing for artist identification: Computerized analysis of Vincent van Gogh's brushstrokes. *IEEE Signal Processing Magazine*, *25*(4), 37–48.

Juola, P. (2006). A prototype for authorship attribution studies. *Foundation and Trends in Information Retrieval*, *1*(3), 233–334.

Juola, P. (2007). Becoming Jack London. *Journal of Quantitative Linguistics*, *14*(2–3), 145–147.

Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management*, *45*(5), 499–512.

Kjell, B., Woods, W. A., & Frieder, O. (1994). Discrimination of authorship using visualization. *Information Processing and Management*, *30*(1), 141–150.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management*, *44*(4), 1448–1466.

Lewis, G. (1999). *The Turkish Language Reform: A Catastrophic Success*. Oxford: Oxford University Press.

Lewis, G. L. (1988). *Turkish Grammar* (2nd ed.). Oxford: Oxford University Press.

Naci, F. (1999). *Yüzyılın 100 Türk Romanı (100 Novels of the Century)*. İstanbul: Adam Yayınları.

Necatigil, B. (1992). *Edebiyatımızda Eserler Sözlüğü (Dictionary of Works in Our Literature)*. İstanbul: Varlık Yayınları.

Ozkar, M., & Lefford, N. (2006). Modal relationships as stylistic features: Examples from Seljuk and Celtic patterns. *Journal of the American Society for Information Science and Technology*, *57*(11), 1551–1560.

Patton, J. M., & Can, F. (2004). A stylometric analysis of Yaşar Kemal's *İnce Memed* tetralogy. *Computers and the Humanities*, *38*(4), 457–467.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, *31*(4), 351–365.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*, 287–313.

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

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 461–485.

Stamou, C. (2007). Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, *23*(2), 181–199.

TDK. (1974). *Türkçe Sözlük (Turkish Dictionary)*. Ankara: TDK (Turkish Language Association).

Woods, M. J. (2001). Spanish word frequency: A historical surprise. *Computers and the Humanities*, *35*(2), 231–236.

APPENDIX

In the following the novels used in the study are listed in the order of first publication year with the following information: author name, the title of the novel (the title in English – if necessary), the original publication year (the publication information for the edition used during OCR).

1. Mehmet Rauf, *Eylül* (*September*), 1901 (Inkilâp Kitabevi, İstanbul, 2003, 4th edition).
2. Hüseyin Rahmi Gürpınar, *Toraman* (*The Young Man*), 1919 (Hilmi Kitabevi, İstanbul, 1948, 2nd edition.
3. Ömer Seyfettin, *Efruz Bey* (*Mr. Efruz*), 1919 (Bilgi Yayınevi, Ankara, 1990).
4. Refik Halit Karay, *İstanbul'un Bir Yüzü* (*An Aspect of İstanbul*), 1920 (Semih Lütfi Kitabevi, İstanbul, 1939, 2nd edition).
5. Reşat Nuri Güntekin, *Çalıkuşu* (*The Autobiography of a Turkish Girl*), 1922, Semih Lûtfi Erciyas Kitabevi, İstanbul, 1942, 6ncı bası, 23üncü bin).
6. Yakup (Yakub) Kadri Karaosmanoğlu, *Nur Baba* (*The Enlightening Father*), 1922 (Remzi Kitabevi, İstanbul, 1939).
7. Halide Edip (Edib) Adıvar, *Kalb Ağrısı* (*The Heart Ache*), 1924 (Halk Kitaphanesi Abdülaziz, publication place is not shown, approximate publication date: 1935).
8. Salâhattin Enis, *Zaniyeler* (*The Adulteresses*), 1924, (Ali Toygar Cumhuriyet Kitabevi, İstanbul 1943, 2nd edition).

9. Halit Ziya Uşaklıgil, *Kırık Hayatlar* (*The Broken Lives*), 1924 (Hilmi Kitabevi, İstanbul, 1944, 2nd edition).

10. Peyami Safa, *Sözde Kızlar* (*The So–called Girls*), 1925 (Sühulet Kitab Evi, İstanbul, 1938, 3rd edition).

11. Mahmut Yesari, *Tipi Dindi!* (*The Blizzard is Over!*), 1933 (Güven Basımevi, Istanbul, 1943, 2nd edition).

12. Memduh Şevket Esendal, *Ayaşlı İle Kiracıları* (*Ayaşlı and his Tenants*), 1934 (Bilgi Yayınevi, Ankara, 2002, 9th edition).

13. Cahit Uçuk, *Dikenli Çit* (*The Barbed Fence*), 1937 (İnkilâp Kitabevi, İstanbul, no publication date is available, 3rd edition, approximate publication date: 1945).

14. Mithat Cemal Kuntay, *Üç İstanbul* (*The Three Periods of İstanbul*), 1938 (Oğlak Yayıncılık ve Reklamcılık Ltd. Şti., İstanbul, 2001, 3rd edition).

15. Abdülhak Şinasi Hisar, *Fahim Bey ve Biz* (*Fahim Bey and Us*), 1941 (Varlık Yayınları, İstanbul, approximate publication date: 1960).

16. Sabahattin Ali, *Kürk Mantolu Madonna* (*The Fur Cloaked Madonna*), 1943 (Varlık Yayınları, İstanbul, 1966, 2nd edition).

17. Kemal Bilbaşar, *Denizin Çağırışı* (*The Call of the Sea*), 1943 (Bilgi Yayınları, Ankara, 1972, 2nd edition).

18. Sait Faik, *Medarı Maişet Motoru* (*The Motorboat of Survival*), 1944 (Bilgi Yayınevi, Ankara, 1970, 4th edition).

19. Ahmet Hamdi Tanpınar, *Huzur* (*A Mind in Peace*), 1949 (Tercüman Kitapçılık, approximate publication date: 1970).

20. Oktay Akbal, *Garipler Sokağı* (*The Alley of Poor People*), 1950 (Cem Yayınevi, İstanbul, 1967, 2nd edition).

21. Orhan Kemal, *Cemile*, 1952.

22. Yaşar Kemal, *İnce Memed* [1] (*Memed, My Hawk*), 1955.

23. Yusuf Atılgan, *Aylak Adam* (*TheLoiterer*), 1959 (Bilgi Yayınları, Ankara, 1974, 2nd edition).

24. Kemal Tahir, *Yorgun Savaşçı* (*The Tired Warrior*), 1965 (Tekin Yayınevi, İstanbul, 2002, 16th edition).

25. Tağrık Buğra, *İbiş' in Rüyası* (*The Dream of İbiş*), 1970 (Ötüken Neşriyat A. Ş., İstanbul, 2000, 11th edition).

26. Fakir Baykurt, *Tırpan* (*The Trepan*), 1970.

27. Çetin Altan, *Büyük Gözaltı* (*The Grand Surveillance*), 1972 (Bilgi Yayınevi, Ankara, 1973, 2nd edition).

28. Oğuz Atay, *Tutunamayanlar* (*The Losers*), 1972 (İletişim Yayınları, İstanbul, 2003, 29th edition of İletişim Yayınları).

29. Adalet Ağaoğlu, *Ölmeye Yatmak* (*Lying Down to Die*), 1973 (Remzi Kitabevi, İstanbul, 1980, 3rd edition).
30. Füruzan, *Kırkyedililer* (*Those Born in '47*), 1974 (Can Yayınları, İstanbul, 1990, 6th edition).
31. Pinar Kür, *Yarın Yarın* (*Tomorrow Tomorrow*), 1976 (Bilgi Yayınevi, Ankara, 1976, 1st edition).
32. Ferid Edgü, *O; Hâkkari'de Bir Mevsim* (*He; A Season in Hâkkari*), 1977 (Yapı Kredi Yayınları, İstanbul, 2002, 10th edition).
33. Selim İleri, *Ölüm İlişkileri* (*Death Relations*), 1979 (Bilgi Yayınları, Ankara, 1979, 1st edition).
34. Orhan Pamuk, *Sessiz Ev* (*The House of Silence*), 1983 (İletişim Yayınları, İstanbul, 1999, 20th edition).
35. Latife Tekin, *Sevgili Arsız Ölüm* (*Dear Shameless Death*), 1983 (Adam Yayınları, İstanbul, 1985, 8th edition).
36. Mehmet Eroğlu, *Issızlığın Ortasında* (*In the Middle of Desolation*), 1984 (Can Yayınları, İstanbul, 1987, 3rd edition).
37. Attilâ İlhan, *Haco Hanım Vay!* (*Haco Hanım Oh!*), 1984 (Bilgi Yayınları, Ankara, 1999, 3rd edition).
38. Kaan Arslanoğlu, *Devrimciler* (*The Revolutionaries*), 1987 (Adam Yayınları, İstanbul, 1997, 2nd edition).
39. Nedim Gürsel, *Boğazkesen Fatih'in Romanı* (*Cut-Throat: The Novel of Mehmet the Conqueror*), 1995 (Can Yayınları, İstanbul, 1998, 6th edition).
40. Ahmet Altan, *Kılıç Yarası Gibi* (*Like a Sword Wound*), 1998 (Can Yayınları, İstanbul, 1998, 30th edition).