*Computers and the Humanities* **38:** 61–82, 2004.



# Change of Writing Style with Time

FAZLI CAN[1] and JON M. PATTON[2]
[1]*Computer Science and Systems Analysis Department, Miami University, Oxford, OH 45056, USA*
*E-mail: canf@muohio.edu*
[2]*Miami Computing and Information Services, Miami University, Oxford, OH 45056, USA*
*E-mail: pattonjm@muohio.edu*

**Abstract.** This study investigates the writing style change of two Turkish authors, Çetin Altan and Yaşar Kemal, in their old and new works using respectively their newspaper columns and novels. The style markers are the frequencies of word lengths in both text and vocabulary, and the rate of usage of most frequent words. For both authors, t-tests and logistic regressions show that the length of the words in new works is significantly longer than that of the old. The principal component analyses graphically illustrate the separation between old and new texts. The works are correctly categorized as old or new with 75 to 100% accuracy and 92% average accuracy using discriminant analysis-based cross validation. The results imply higher time gap may have positive impact in separation and categorization. For Altan a regression analysis demonstrates a decrease in average word length as the age of his column increases. One interesting observation is that for one word each author has similar preference changes over time.

**Key words:** agglutinative languages, authorship attribution, statistical analysis, stylochoronometry, stylometry, Turkish

## 1. Introduction

Researchers challenge themselves in stylometry by analyzing writing styles of authors using objective measures. For this purpose various style markers (measurable attributes) are defined and their occurrence patterns in the text of interest are examined using statistical methods. These patterns are used to resolve stylometric problems, such as authorship attribution and stylochronometry (i.e., assigning date to work). The objectives of stylometry are similar to data mining, and clustering (unsupervised learning). For example, data mining tries to discover understandable and valid patterns hidden in data (e.g., credit card transactions) that can be interesting and useful (Fayyad *et al.*, 1996). In clustering, objects (patterns, documents, etc.) are grouped into sets containing similar items (Jain *et al.*, 1999). In stylometry we have a supervised classification problem, since we are provided with a set of pre-classified objects and try to categorize a newly encountered object into these existing sets as in the case of information filtering (Foltz and Dumais, 1992).

Using statistical patterns as measures of style may be criticized and regarded as dry by literary scholars; however, style markers are well suited to statistical analysis, and it is shown that the statistical approach works successfully for

numerous cases (Forsyth and Holmes, 1996). Also experiments show that objective measures based on style markers can match the literary critical remarks (Whissell, 1994).

In this study we investigate the style change of two Turkish authors, Çetin Altan and Yaşar Kemal, in their old and new works. In Altan's case old and new works are, respectively, defined as the columns published in the ten-year period 1960–1969, and the year 2000. In Kemal's case as an old work we use his 1971 novel "Bin Boğalar Efsanesi," (Kemal, 1971) and as a new work we use his 1998 novel "Fırat Suyu Kan Akıyor Baksana" (Kemal, 1998). The objective measures of this study show that there is a significant difference between the old and new works of these authors within the context of the chosen works. The change can be attributed to a natural theme shift in the new and old works or a conscious style choice made by the writers. For example, Kemal indicated that in some different novels he used completely different languages because of their locales (Naci, 1999, p. 29). However, since what we use in style analysis is statistical patterns, which are unconsciously chosen by the authors, we hypothesize that style changes as suggested by the style markers are due to the time gap between the works.

Altan and Kemal, whose works we study in this article, are among the most well known writers in Turkey. For example, Naci, in his collection of the reviews of one hundred Turkish novels of the century, included one work of Altan, and five works of Kemal, and all together sixty-two different writers (Naci, 1999).

Writers would have more chance of (unconsciously) controlling their word lengths in agglutinative languages such as Turkish. Therefore, we hypothesize that word length occurrence frequency information is a good measure to use in stylometric investigations in such languages. In this study first we compared the average word length between the old and new works of fixed size blocks for each author using a t-test. The results indicated that the average word length of the new works (i.e., blocks) was significantly larger than that of the old for both authors. This gave us the motivation to perform a regression analysis using Altan's data to show that as the age of the work increases the average word length decreases. After this, for each author, a series of logistic regressions were conducted to test for differences of token and type length frequencies between the old and new works. Then we compared the rate of usage of most frequent words between old and new for both authors, and performed a principal component analysis to graphically illustrate the differences between the old and new works. Finally, we conducted a stepwise discriminant analysis to determine the best discriminators and then used cross validation to determine the categorization success rate using these discriminators.

This article is the first published stylometry study on modern Turkish literature. Our investigation shows that the three style markers and the multivariate techniques we used provide outstanding tools for separating and distinguishing the old and new works of these Turkish authors. This indicates that the same markers and techniques are promising in other stylometric studies in Turkish.

The rest of the paper is organized as follows. In Section 2 we briefly review previous stylometry work. In Section 3 we present the characteristics of the Turkish language in terms of its morphology. The description of the test data and experimental design of the study is given in Section 4. The experimental results and their discussion are presented in Section 5, and finally the conclusions and future research pointers are provided in Section 6.

## 2. Previous Work

For a long time various statistical markers have been used to investigate the characteristics of artifacts in the humanities and fine arts (Sedelow, 1970). A detailed overview of the stylometry studies in literature within a historical perspective can be seen in Holmes (1994). It gives a critical review of numerous style markers. It also reviews works on the statistical analysis of change of style with time. A solid critique of many authorship studies is provided in Rudman (1998). The coverage of these studies is extensive and they come with broad reference lists.

In stylometry studies about 1,000 style markers have been identified (Rudman, 1998). One of the oldest style markers is word length. For example, in 1901 Mendenhall published a well-known study using the word length frequencies and concluded that due to their style difference it was unlikely that Bacon could have written works attributed to Shakespeare. A later work showed that this conclusion could be false, since the style difference of these two authors could be due to different types of works used for comparison (prose of the former and verse of the latter author). This conclusion was based on the fact that in the writings of Sir Philip Sidney, a contemporary of Bacon and Shakespeare, the differences in word lengths between his prose and verse are very close to the differences found between Bacon's prose and Shakespeare's verse (Williams, 1975). However, Holmes (1985) does not give a positive recommendation for the use of word length frequencies in authorship attribution by pointing out the characteristics of Zipf's first law (Zipf, 1932). Another work (Tallentire, 1972) discusses the difficulty of using word length frequencies in authorship studies. It is hard to find similar studies on word length as a style marker in agglutinative languages such as Turkish. In this study we use this style marker and show that it can be utilized successfully as suggested by its comparable (but not as good) results with those of most frequent word occurrence statistics. Our approach of using various style markers also matches Rudman's point that many different style markers need to be taken into account jointly (Rudman, 1998).

An extensively used style marker is the frequency count of "context free" words (or similarly "most frequent words," and "function words"). The paper (Forsyth and Holmes, 1996) studies the use of five style markers (letters, most frequent words, most frequent digrams, and two methods of most frequent substring selection approaches) in ten stylometry problems (such as authorship, chronology, subject matter, etc.) with various levels of success. Another study on authorship attribu-

tion (Baayen *et al.*, 1996) compares the discriminatory power of frequencies of syntactic rewrite rules, lexical methods based on some measures of vocabulary richness, and the frequencies of the most frequent fifty words. The study states that frequencies of syntactic constructs lead to a higher classification accuracy. The work also states that syntax based methods are computationally expensive since they require syntactically annotated corpora. In this study we utilize the usage rate of most frequent words as a style marker.

For analyzing the occurrence patterns of style markers various statistical methods are used. One popular technique in stylometric studies is principal component analysis. It can be easily appreciated using plots usually in two dimensions. For example, the Binongo and Smith (1999) study illustrates the use of the principal component analysis technique using occurrence frequency counts of two words, explains its intuition, and then uses it with several words in the authorship study of one of Shakespeare's romances. We used this statistical technique to visually see the separation between old and new works using our style markers.

Another statistical technique we used in this study is discriminant analysis that is also used in the literature for various purposes. For example, a recent work (Stamatatos *et al.*, 2001) uses discriminant analysis and attacks the authorship detection problem using low-level measures (e.g., sentence length, punctuation mark count, etc.), syntax-based measures (e.g., noun phrase count, verb phrase count, etc.), and a set of style markers obtained by a natural language processing tool (e.g., percentage of rare or foreign words, a measure that indicates the morphological ambiguity, etc.). Additionally, they also use frequencies of most frequent words. The study is especially interesting due to its use of rich combination of style markers. The Baayen *et al.* (1996) paper mentioned before applies discriminant analysis to determine authorship attribution using syntax-based methods. In Holmes and Singh's (1996) paper, a discriminant analysis is conducted to determine what measures of linguistic ability best discriminate aphasic patients from the normal person. The study reported in Holmes and Forsyth (1995) uses discriminant analysis to determine which vocabulary richness measures best discriminated between the papers written by Alexander Hamilton and those by James Madison. The study reported in Martindale and Tuffin (1996) uses it to find differences between Homer's Illiad and Odyssey.

Another multivariate technique used was logistic regression. This technique is useful for studying curvilinear relationships between a binary response variable and one or more predictor variables. The study reported in Kessler *et al.* (1997) employs logistic regression in genre detection and compares its performance with a neural network approach. In our work, this method appears appropriate when the response variable is the classification of a work being "old" or "new" and the predictors are frequencies of word lengths or of the usage of "most frequent" words.

The only stylometric work (Private Communication, Gökhan Tür, 2001, unpublished work), that we were able to find on the Turkish language, studies authorship attribution using the unigram language model (Ney *et al.*, 1994) which is based on

*Table I.* Turkish alphabet

| Vowels | Consonants |
|---|---|
| a, e, ı, i, o, ö, u, ü | b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z |

the occurrence probabilities of words. In Tür's study, when an unseen text is given, the system chooses the writer whose corresponding language model maximizes the probability of authorship for that piece of text. The approach gives more than 90% correct author detection results. The study shows that stemming decreases the accuracy since it eliminates important stylistic information.

## 3. Turkish Language Morphology

Turkish belongs to the Altaic branch of the Ural-Altaic family of languages. It is a free constituent order language, i.e., according to text flow and discourse context at certain phrase levels, its constituents can change order (Hakkani-Tür, 2000). Turkish is an agglutinative language similar to Finnish, and Hungarian. Such languages carry syntactic relations between words or concepts through discrete suffixes, and they have complex word structures. Turkish words are constructed using inflectional and derivation suffixes. The study of Turkish morphology as a computation problem can be found in Köksal (1973) and Solak and Oflazer (1993). A two-level (lexical and surface) morphological description of Turkish word structure is studied in Oflazer (1994). Statistical modeling and its use in morphological disambiguation, spelling correction, and speech recognition are studied in Hakkani-Tür (2000).

The Turkish language alphabet, in its current orthography, is based on Latin characters and has 29 letters (see Table I). It contains eight vowels, and 21 consonants. In some words borrowed from other languages, such as Arabic and Persian, the vowels "a", "i", and "u" are made longer by using the character ^ on top of them. In modern spelling this approach is rarely used.

In agglutinative languages it is possible to have words that would be translated into a complete sentence in languages such as English. One common example, which is used in various forms by grammar students in Turkey, is the following Turkish word-sentence.[1]

"ÇEKOSLOVAKYALILAŞTIRAMADIKLARIMIZDANSINIZ."

This word can be divided into morphemes as follows (the + signs indicate morpheme boundaries):

"ÇEKOSLOVAKYA+LI+LAŞ+TIR+AMA+DIK+LAR+IMIZ+DAN+SINIZ"

It can be translated into English as "You are one of those whom we were unable to convert to a Czechoslovakian." (Most probably the country name "Czechoslovakia" has been chosen because of its relatively long length in addition to factual suggestions of the word.) However, note that this is an exaggerated example reflecting the agglutinative nature of the language.

To better reflect the characteristics of the language some numbers regarding its morphological structure are given in the following. In Turkish the number of possible word formations obtained by suffixing one morpheme to a "noun" type stem is 33. By adding two and three morphemes to a noun type of stem one can obtain 490 and 4,825 different words, respectively. For an "adjective" type word stem the respective numbers if we add one, two, and three morphemes are 32, 478, and 4,789. For "verb" type word stems the number of possible word formations, respectively, are 46, 895, and 11,313 (Hakkani-Tür, 2000, p. 31).

## 4. Test Data and Experimental Design

In this study an individual text word, *token*, is defined as a continuous string of word characters. A distinct word, *type*, is defined as a set of identical individual words. The term *vocabulary (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 letters are not case sensitive. The "word characters" are the Turkish alphabet letters, the versions of "a" and "I" with a ^ on top of them (in the corpora there was no occurrence of u with ^ on top of it), the letters (q, x, w), the numerals 0 to 9, and the apostrophe sign. The letters q, x, and w are included since in Altan's writings they are used in foreign proper nouns. As expected, such words constitute an insignificant portion of the data sets. The minimum word length is defined as two. When we collect word length frequencies we consider both tokens and types.

### 4.1. Description of old and new works

For the creation of our test cases we downloaded Altan's columns from the web site of the Sabah newspaper. In this newspaper on Mondays Altan published a work written several years ago. We compiled all such works from Sabah, published from January 1, 1997 up to and including February 19, 2001 (Altan, 1997–2001). (Note that the Sabah newspaper issues are available on the web beginning with the year 1997.) We were able to compile 201 unique past works with a publication time range of 46 years from 1945 to 1991. The reader may refer to Table II for the number of columns obtained for each of those years.

For Altan we were looking for a period of time which is far-off from today and that would yield the highest number of columns in the experiments. Under these constraints 1960 to 1969 is the best period of time providing 99 columns with 41,783 words (accordingly the average column size is 422 words). For his new works, we used columns published in 2000 beginning with the first day of the year.

*Table II.* Column per year information for Altan's works

| Year | No. of Col. | Year | No. of Col. | Year | No. of Col. |
|---|---|---|---|---|---|
| 1945 | 1 | 1967 | 15 | 1982 | 14 |
| 1947 | 1 | 1968 | 6 | 1983 | 11 |
| 1949 | 2 | 1969 | 5 | 1984 | 11 |
| 1953 | 1 | 1971 | 1 | 1985 | 10 |
| 1954 | 1 | 1973 | 2 | 1986 | 1 |
| 1960 | 2 | 1975 | 1 | 1987 | 6 |
| 1961 | 12 | 1976 | 4 | 1988 | 8 |
| 1962 | 15 | 1977 | 9 | 1989 | 1 |
| 1963 | 12 | 1978 | 1 | 1990 | 1 |
| 1964 | 16 | 1979 | 1 | 1991 | 1 |
| 1965 | 11 | 1980 | 9 | — | — |
| 1966 | 5 | 1981 | 4 | — | — |

In Kemal's case the old work used was his 1971 novel "Bin Boğalar Efsanesi," (Kemal, 1971) and the new work used was his 1998 novel "Fırat Suyu Kan Akıyor Baksana" (Kemal, 1998). The publication year separation of these two works is 27 years. For his new work we used the version that was originally published in the Milliyet newspaper web site (http//www.milliyet.com) between the dates November 23, 1997 and January 1, 1998. This version of the novel is slightly shorter and towards the end some segments are presented in slightly different order than its published book version.

### 4.2. Experimental design and characteristics of data sets

For principal component analysis, discriminant analysis, t-tests, and regressions, we needed observations based on fixed size text blocks, i.e., blocks containing the same number of words. After reviewing the literature we decided that 2,500 words is an appropriate block size to be used (Binongo and Smith, 1999, p. 460; Forsyth and Holmes, 1996, p. 164; Baayen *et al.*, 1996, p. 122). In block generation Altan's columns are placed one after the other, and the obtained text has been divided at every 2,500th word. In Kemals' novels the first 2,500 words constituted the first block and so on. The smallest amount of data we have is for the old works of Altan. That provided us with 16 blocks. Since we wanted to have a balanced experimental design, the size of the old works of Altan determined the number of blocks to be used in the experiments.

For comparison purposes and to see the characteristics of the data sets, information is given in Table III for words forming the old and new blocks of Altan, the

*Table III.* No. of tokens, types and their length information for some Turkish and English text

| Text | No. of tokens (N) | No. of types (V) | Avg. token length | Avg. type length |
|---|---|---|---|---|
| Çetin Altan, old col.s (16 blocks) | 40,000 | 14,926 | 6.25 | 8.10 |
| Çetin Altan, new col.s (16 blocks) | 40,000 | 14,459 | 6.52 | 8.51 |
| Çetin Altan, old new cols. together | 80,000 | 25,392 | 6.39 | 8.54 |
| Yaşar Kemal, Bin Boğalar Efsanesi | 66,969 | 15,491 | 5.90 | 8.04 |
| Yaşar Kemal, Fırat Suyu Kan Akıyor | 73,043 | 16,450 | 5.99 | 8.18 |
| Yaşar Kemal, both novels together | 140,012 | 25,843 | 5.95 | 8.34 |
| Ahmet Hamdi Tanpınar, Huzur | 97,748 | 23,407 | 6.21 | 8.54 |
| Turkish Newspapers | 709,121 | 89,103 | 6.52 | 9.28 |
| Kucera and Francis | 1,014,232 | 50,406 | 4.74 | 8.13 |

two novels of Kemal, and a novel by Ahmet Hamdi Tanpınar, "Huzur," originally published in 1949 (Tanpınar, 1982). In Kemal's case, rather than giving the old and new block related information (as we did for Altan), we preferred to give word information for the complete novels since they are integral units. The Tanpınar case is given to provide a comparison with another Turkish author whose work is considered significant by literary critics (Naci, 1999, p. 245). We also performed an analysis of news articles and some columns compiled from Turkish newspapers and provided the numbers in the same table (TNP, 2000). To see things in a different context we also provided numbers on American English taken from (Kucera and Francis, 1967, pp. 365–366). Their work provides the average token and type lengths for text of different genres. In their study the word definition is slightly different than ours; however, if we had used their definition it would have imposed an insignificant effect on our numbers due to the nature of our data sets.

The numbers show that Altan's token lengths are longer than those of Kemal's and Tanpınar's. The rows for Altan's both work types combined and Kemal's both novels combined indicate that the average type length increases as the number of tokens increases. This is expected, since as we go through a text (whether Turkish or English) we get the more frequent word types first during the early sampling process, and less frequent (i.e., less probable) word types, later in the sampling process, add to this the well-known correlation of word frequency and length, it follows that we sample more short word types initially, and as we keep accumulating tokens, the new word types we see tend to be longer (Baayen, 2001, p. 197). (Of course, this type length increase would reach a steady state value after considering enough number of tokens due to almost no increase in vocabulary size (Heaps, 1978, p. 206).)

Average token and type lengths for all Turkish writers and the newspapers is longer than that of English words. For example, the average token and type lengths in Turkish newspapers is 38% and 14% longer than that of the average English

*Table IV.* Altan's most frequent words used in the analysis (in descending order of frequency when both old and new blocks are combined)

| Turk. | bir | ve | de | da | ne | bu | için | daha | gibi | çok | ki | sonra | kadar | hiç |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Engl. | a<br>an<br>one | and | too | too | neither<br>nor,<br>what | this | for | more | like | very | that<br>which<br>who | later | until<br>till | nothing |

*Table V.* Kemal's most frequent words used in the analysis (in descending order of frequency when both novels are combined)

| Turk. | bir | de | bu | da | ne | gibi | dedi | kadar | çok | daha | ben | sonra | diye | ki | her |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Engl. | a<br>an<br>one | too | this | too | neither<br>nor.<br>what | like | said | until<br>till | very | more | I | later | that* | that<br>which<br>who | every |

*An example "Eğleniriz diye gittik." in English it means "We went hoping that we would amuse ourselves." For other examples see Redhouse (1979, p. 305).

token (6.52 vs. 4.74) and type (9.28 vs. 8.13) respectively. In other words, the difference in terms of type lengths is not as dramatic as token lengths. This can be explained by the fact that the multiple occurrences of long words are discarded when they are inserted into the vocabulary. Longer word lengths in Turkish are due to the agglutinative nature of the language.

The average type and token lengths of Turkish newspapers are longer than those of the Turkish authors listed in Table III with the exception of Altan's average token length of his new columns. This may be due to the factual content of the newspaper collection; the descriptive nature of such content may require the usage of longer words.

For Altan's so called context-free words we considered the 25 most frequent words (top-25) of the old and new blocks (40,000 words each) and used the 14 common words of these two lists. The selected words were in the top-15 list of all columns of Altan published in 1999 and 2000 containing a total of 264,184 tokens excluding quotations from poems. The selected 14 words of Altan with their closest English translation are given in Table IV in descending order of frequency in the old and new blocks combined.

For the selection of Kemal's so called context-free words we considered the top-20 words from both novels and used the common words. This approach provided us with 15 different words and they are listed in Table V. When we combine the two novels, the selected 15 words appear in the top-16 of the combined works.

Some observations about the selected most frequent words are in order. From Table IV and V Altan and Kemal have 11 common most frequent words. The words "ben," "dedi," "diye" of Kemal are more appropriate for story telling, and

that is why they do not appear in Altan's list. Twelve words of Table IV, Altan's words, appear in the top-15 and top-18 of Kemal's combined works and the Turkish newspapers, respectively. Although the word "ve" has been used quite infrequently in Kemal's works (when both novels are combined the word "ve" is the 127th most frequent word along with one other word), it is in Altan's list. Also "ve" is the second and third most frequent word in the newspaper collection and Tanpınar's "Huzur" respectively.

## 5. Experimental Results and Discussion

In this section we first compared the average token and type lengths between the old and new works for each author using a t-test (or a one way analysis of variance). The results indicated that the average token and type length of the new works was significantly larger than that of the old for both authors. This gave us the motivation to perform a regression analysis using Altan's data to show that, as the age of the work increases, the average token length decreases. Then for each author, a series of logistic regressions were conducted to test for differences of token and type length frequencies between the old and new works. Then we compared the rate of usage (i.e., usage frequency) of most frequent words between old and new for both authors. We next performed a principal component analysis using our all style markers and then created scatterplots of principal component scores for each data corresponding to a text block. This is to graphically illustrate the differences between the old and new works by looking at their principal component scores. Finally we conducted a stepwise discriminant analysis to determine the best discriminators and then used cross validation to determine the success rate using these discriminators. All of these analyses were conducted using the SAS for Windows software, Version 8.

### 5.1. Comparing token and type lengths in old and new works

#### 5.1.1. *Exploratory Comparison of Token and Type Lengths in Old and New Works*

A one way analysis of variance for each author was conducted on the blocks of the two work types to determine if there are differences in the average token and type lengths between the two. The results are summarized in Tables VI and VII for Altan and Kemal.

For both token and type, the average word lengths for each author were significantly larger for the new writings compared to the old. However, for Kemal the difference was not as dramatic as that for Altan.

The above results motivated us to perform a regression analysis to determine whether a relationship exists between average token length and age of the work. We could only conduct the analysis using Altan's columns as a data source since they were written during several time points. The average token length in Altan's

*Table VI.* Comparison of token and type lengths between old and new works of Altan

| Word Type | Average word length of old works | Average word length of new works | Pooled standard deviation | Test statistic | P-value |
|---|---|---|---|---|---|
| Token | 6.25 | 6.52 | 0.115 | 6.50 | <0.0001 |
| Type | 7.27 | 7.62 | 0.120 | 8.11 | <0.0001 |

*Table VII.* Comparison of Token and Type Lengths Between Old and New Works of Kemal

| Word type | Average word length of old works | Average word length of new works | Pooled standard deviation | Test statistic | P-value |
|---|---|---|---|---|---|
| Token | 5.92 | 6.03 | 0.138 | 2.12 | 0.0427 |
| Type | 6.96 | 7.09 | 0.120 | 2.95 | 0.0030 |

columns during a given year was chosen as the response variable and the age of the columns as the independent variable. For the regression analysis we used all the works stated in Table II and the 16 columns for each year between 1997 and 2001; i.e., 80 columns in addition to the ones of Table II, and therefore a total of 281 columns. Note that sixteen is the maximum number of columns we have for the years earlier than 1997 (see Table II). For years with several columns the mean token length, averaged over all the columns for that year, was used as the dependent data value. This results in 39 observations (i.e., different years) used in this analysis. Figure 1 is a scatterplot of average token length versus year for the works of Altan. There appears to be a linear trend.

The regression results indicated a very strong relationship between average token length and age (F(1,37) = 11.78, Prob-value = 0.0015). The $R^2$ statistic was .2415 indicating that 24.15% of the total variance of average token length per year about the overall average token length can be explained by this regression model. There may be other factors affecting average text word length but the age of the "work" (i.e., column, or average column if we have many columns for that year) is an important factor. The prediction equation is given by

$$\text{Average token length} = 6.5031 - 0.00665 * \text{age of work}$$

Thus as the age of the work increases by one year, the mean of the average token length would decrease by approximately 0.00665, where age is defined as 2001 minus the publication year of the work.

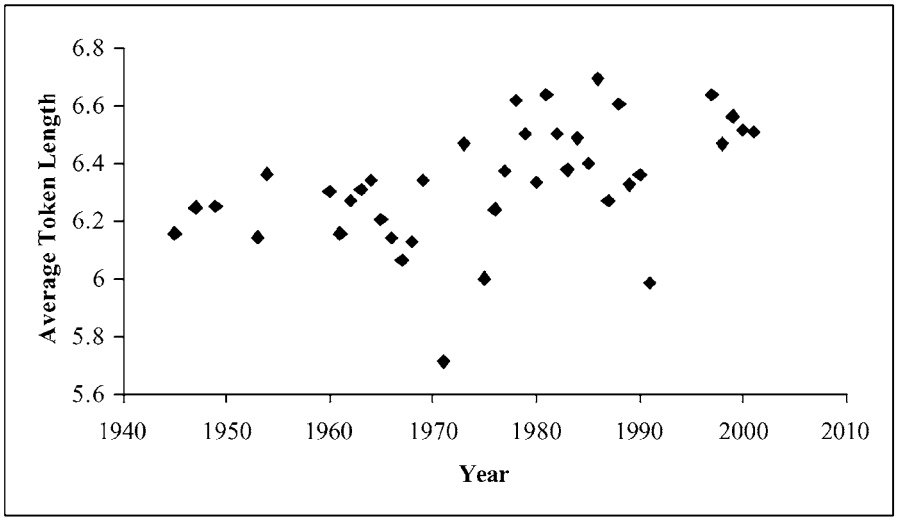


*Figure 1.* Average token length for each year (for the works of Altan).

### 5.1.2. *Detailed Comparison of Token and Type Lengths in Old and New Works*

5.1.2.1. *Comparison of token and type lengths in old and new works of Altan.* Logistic regressions were then conducted to determine whether significant differences in the frequencies of token and type lengths existed between the two work types. From the above results, it would appear that the more recent writings would have a larger percentage of longer words than the older writings. The binary response variable used in each logistic regression was the classification of the work type (new, old). The independent variable was the frequency of tokens or types of a certain length. A regression was conducted for each length starting at 2 letters and going through 19 and done separately for tokens and types. The results are given in Appendix Table App. I.

The first two columns contain the mean and standard deviation of the number of occurrences for both token and type for each word length. The next column contains an odds ratio, which is a measure of association. The odds ratio compares the odds of finding a word belonging to an old work to the odds of belonging to a new work when the word of a certain length is chosen at random. An odds ratio less than one indicate that such a word is more likely to come from a new work whereas a ratio greater than one indicates a greater likelihood that it is from an old. The last column is the observed significance level of the odds ratio based on Wald's Chi-Square statistic

The results indicate that for word lengths of 8 or less for both token and type (except for tokens of length 2), there were better odds of finding the word in an old work than in a new. For word lengths of 9 or greater, there was greater likelihood of finding the word in a new work and for word lengths of 11 or greater, these

differences were significant at the 0.05 significance level. This was true for both token and type.

5.1.2.2. *Comparison of token and type lengths in old and new works of Kemal*. As in Altan's case, a similar analysis was conducted for differences in the frequencies of both the token and type lengths of Kemal. The results, which are given in Appendix Table App. II, indicate that for tokens of lengths 3, 4, 5, and 6 and for types of length 5 or less there were a larger average number of occurrences in the old works. This was also true for tokens and types of length 16. For types of length 2, there was significantly more usage in the old writings (since the observed significance levels were less than 0.05). For types of length 3, 10, 11, and 12, the usage in the new writings was somewhat significantly larger (observed significance level was less than 0.10).

Comparing the tokens, there was a significantly higher frequency of 4 character tokens in the older writings. For tokens of length 2, 10, and 12 the usage in the new writings was somewhat significantly larger (observed significance level was less than 0.10).

## 5.2. Comparison of most frequent word counts between old and new works

We next determined whether there were differences in the usage of most frequent words between the old and the new works. For both Altan and Kemal a series of logistic regressions were conducted to determine if the frequencies of usage of the words given in Tables IV and V differed between the old and new works. Table App. III in the Appendix summarizes the results.

### 5.2.1. *Comparison of Most Frequent Word Counts between Old and New Works of Altan*

In Altan's case from the table, the words "bu," "çok," "da," "de," and "gibi," showed significant different frequency counts between the old and new writings. The words "bu," and "gibi," were used more frequently in the old writings; whereas, the words "çok," "da," and "de" were used more in the new.

### 5.2.2. *Comparison of Most Frequent Word Counts between Old And New Works of Kemal*

The reader may again refer to Table App. III in the Appendix for a summary of the results. From the table, the words "daha," "dedi," "gibi," and "kadar" showed significant different frequency counts between the old and new works. The words "dedi" and "gibi" were used more frequently in the old blocks; whereas, the words "daha," and "kadar" were used more in the new.

Between these two authors we also have the interesting observation that they have a similar preference change over time for the word "gibi." It is used less frequently in the new works for each author.

### 5.3. Principal component analysis results

A series of principal component analysis was conducted on the works of each author. The purpose is to transform each of the three sets of related variables: the frequencies of both token and type lengths, and the rate of usage of the most frequent words into a set of uncorrelated variables called principal components. The data used in this analysis, as well as the discriminant analysis presented below, consists of the 16 old and 16 new works from each author (please also refer to Section 4.2). The frequencies of each token and type length from 2 thru 19 characters and the counts of each of the most frequent words were the variables in the respective analysis.

The plots of the first two principal components in the six analyses are presented in Figure 2. These graphs illustrate a better separation of the old and new works for Altan than Kemal. For both authors the separation appears from best to worst when the two principal components are derived from the rate of usage of most frequent words, type lengths, and token lengths respectively.

### 5.4. Discriminant analysis results

A stepwise discriminant analysis was conducted to determine what token length frequencies provide the best separation between the work types. Using these length frequencies, an additional discriminant analysis was conducted to determine the percentage of blocks correctly classified using cross-validation. In cross validation each block in turn is excluded from the rest of the blocks in the derivation of linear discriminant functions employed for classifying each block as old or new. Then the excluded block is classified by these linear discriminant functions. This eliminates bias from the classification procedure. A similar set of analyses was conducted for type length frequencies and rate of usage of most frequent words.

For Altan the best discriminators among the token length frequencies are those of the (in decreasing order of discrimination power) 4, 8, 9, 11, 14, 15, and 18 character words. The best type length frequency discriminators were those of the 3, 4, 8, 9, 11, 14, and 15 character words. The best discriminators among the most frequent words are "da," "de," "çok," "bu," "için," and "sonra."

In the Kemal case the best discriminator among the token length frequencies were the words containing four characters. The set of type length frequencies used as discriminators were those of the 2, 4, 7, 11, and 18 character words. The rate of usage of the most frequent words "dedi," "kadar," "da," "sonra," "ne," and "daha" were considered the best discriminators.

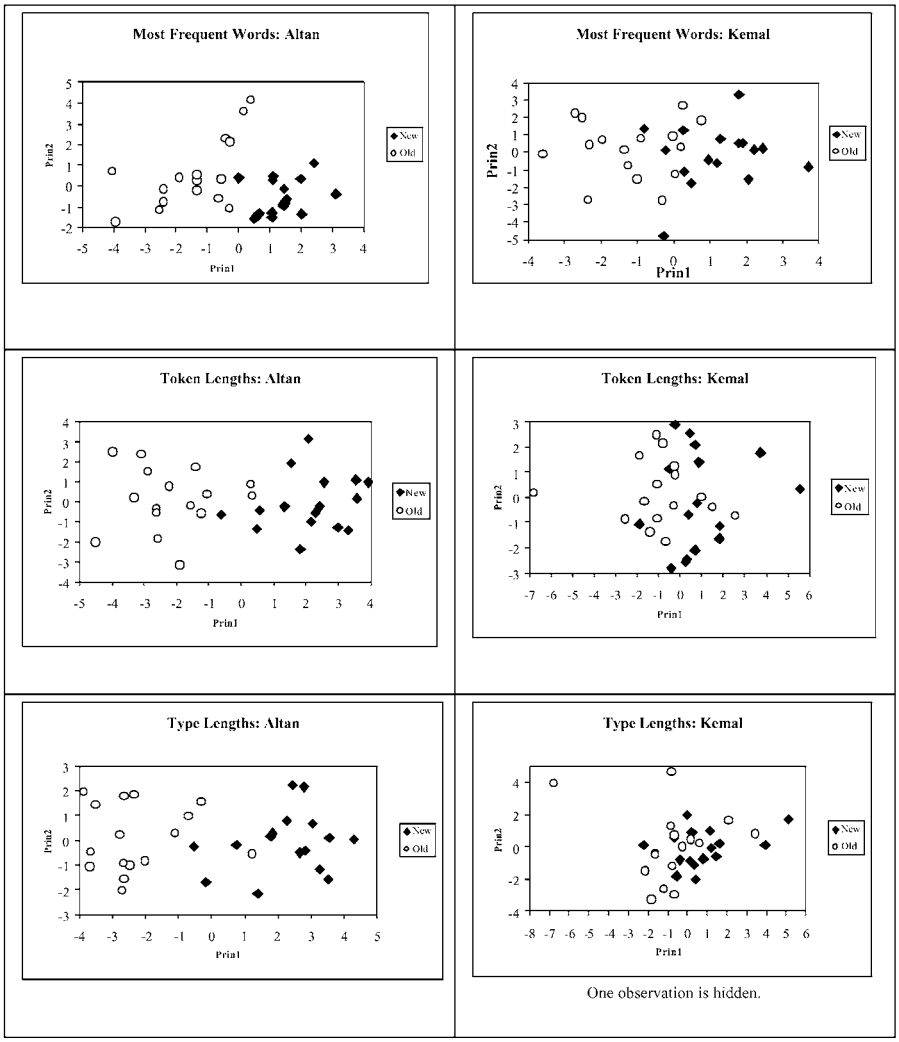


*Figure 2.* Principal component plots for Altan and Kemal based on frequency information.

Table VIII gives the success rate for our discriminant analysis experiments using cross validation. For example, the old works of Altan are classified with the percentage success rate of 93.75, 100.00, and 100.00 using token length frequency, type length frequency, and most frequent word usage rates respectively. The average success rate for Altan and Kemal are, respectively, 98.96 and 84.38%, and the overall average success rate for both authors is 91.67%. Note that the time gap between Altan's old and new works is much higher than Kemal's. The better results with Altan's writings may imply that a higher time gap between works increases the success rate and this can be explained by increased style change with time.

*Table VIII.* Success rate of cross validation in percentages (No. of successful cases)

| Data type | Work type | | | |
|---|---|---|---|---|
| | Altan works | | Kemal works | |
| | New | Old | New | Old |
| Token length | 100.00 | 93.75* | 81.25 | 75.00 |
| Frequencies | (16) | (15) | (13) | (12) |
| Type Length | 100.00 | 100.00 | 75.00 | 81.25 |
| Frequencies | (16) | (16) | (12) | (13) |
| Most Frequent | 100.00 | 100.00 | 93.75 | 100.00 |
| Word Use Rate | (16) | (16) | (15) | (16) |

*93.75% (15) of the 16 old blocks are successfully classified as old.

## 6. Conclusions

In this study we compare the old and new writing style of Çetin Altan and Yaşar Kemal, by employing three style markers: the frequencies of token lengths, type lengths, and most frequent words. Writers would have more chance of (unconsciously) controlling their word lengths in agglutinative languages such as Turkish. Therefore, intuitively word length occurrence frequency information is a good measure to use in stylometric investigations in such languages. We used the newspaper columns of Altan and the novels of Kemal. The separation of the old and new works of Altan is 30 to 39 years, and of Kemal is 27 years.

Principal component and discriminant analyses were conducted to determine the separation between old and new works. The principal component analyses graphically illustrate their separation. The principal components based on most frequent words graphically illustrate the best separation between old and new. In general the separation of Altan's old and new works are better than that of Kemal's. Using discriminant analysis, each text block is classified according to work type. In these tests the overall average success rate is 91.67%, and the average success rates for Altan and Kemal are 98.96 and 84.38%, respectively. Again the success rate of both type and token length frequencies is slightly less than that of most frequent words results. Similar observations have been stated in the literature. For example, Holmes (1985) does not give a positive recommendation for the use of word length frequencies in authorship attribution. Our approach of using various style markers matches Rudman's point which states that many different style markers need to be taken into account jointly (Rudman, 1998). Better separation and higher success in Altan's principal component and discriminant analyses experiments may be attributed to more style change due to a higher time gap between his works.

Our statistical analysis indicates strong relationships between average token length and the age of the work. The regression analysis based on Altan's 281 distinct newspaper columns published between 1945 and 2001 show that as the age of the work increases the average token length decreases. For both authors, t-tests and logistic regressions show that the length of the words in new works is significantly longer than that of the old. The change in the word length may be explained as an effect of time, i.e., in Turkish, as time passes, writers may use longer words as a result of their higher mastery of the language. However, this could as well indicate a change in the Turkish language as a whole during the same time period. Given that there is no accessible diachronic corpus of Turkish or qualitative studies of language change covering a similar period this needs much effort to adequately investigate. The shorter word lengths observed in Kemal's work could be attributed to his use of the inner thoughts of and conversations among his novel characters that are usually countryside people. Furthermore, in both novels used in this study, Kemal as the storyteller frequently uses the languages of his characters.

Between these two authors we also have the interesting observation that they have similar preference changes over time for the word "gibi." It is used less frequently in the new works for each author. Further research needs to be conducted to determine whether other authors have similar preferences for this word.

In terms of future research directions an interesting possibility is the study of Kemal's four "İnce Memed" novels, which were published in a thirty-two year period between 1955 and 1987. For example, according to Naci (1999, pp. 29, 404, 411) Kemal used the same style in all volumes of his novel İnce Memed while using different style or styles in the other novels he published in the same period of time. However, there are different ideas on this "the same style" view (Çiftlikçi, 1997, p. 184; Private Communication, Süha Oğuzertem, 2002). Studying the effect of time on these novels in a stylometric dimension sounds interesting.

The three style markers used in this study and the multivariate techniques provide outstanding tools for separating and distinguishing the old and new works of these Turkish authors. By no means are the style markers used in this study the only ones that can be used to measure the style change with time. It may be interesting to see the results based on some other style markers. In the use of some of these markers, such as "vocabulary richness," morphological structure of the Turkish language may need to be addressed and will challenge researchers.

**Acknowledgements**

We would like to thank the anonymous referees for their useful comments that made the presentation and the content of the paper better. We also acknowledge various comments of Dr. Süha Oğuzertem of the Turkish Literature Department of Bilkent University for the first version of this paper. The Turkish newspapers collection we used in Table III is a courtesy of Dr. Kemal Oflazer of Sabancı University.

*Appendix I.* Altan logistic regression results comparing token and type lengths between old and new works

| | | Tokens | | | | Types | | | |
|---|---|---|---|---|---|---|---|---|---|
| Word len. | Work type | Average number of of occur. | Stand. dev. | Odds ratio | Prob > Chi-square | Average number of occur. | Stand. dev. | Odds ratio | Prob > Chi-square |
| 2 | Old | 183.31 | 32.22 | 0.981 | 0.202 | 23.81 | 2.29 | 1.042 | 0.8034 |
| | New | 195.13 | 16.22 | | | 23.62 | 2.09 | | |
| 3 | Old | 279.69 | 34.14 | 1.035 | 0.0263 | 65.88 | 7.19 | 1.068 | 0.254 |
| | New | 252.25 | 23.90 | | | 63.19 | 5.96 | | |
| 4 | Old | 259.13 | 19.03 | 1.052 | 0.0394 | 120.56 | 10.81 | 1.088 | 0.0504 |
| | New | 244.94 | 15.10 | | | 112.25 | 10.27 | | |
| 5 | Old | 410.63 | 28.51 | 1.028 | 0.0896 | 231.88 | 15.16 | 1.104 | 0.0134 |
| | New | 394.5 | 21.04 | | | 217.38 | 9.25 | | |
| 6 | Old | 298.19 | 27.01 | 1.028 | 0.0847 | 213.5 | 16.55 | 1.075 | 0.0084 |
| | New | 282.38 | 21.69 | | | 191.6 | 18.80 | | |
| 7 | Old | 296.19 | 21.75 | 1.052 | 0.0371 | 237.44 | 19.11 | 1.188 | 0.0099 |
| | New | 280.00 | 15.23 | | | 210.81 | 9.16 | | |
| 8 | Old | 251.63 | 15.30 | 1.160 | 0.0063 | 207.38 | 11.93 | 2.173 | 0.0001 |
| | New | 224.56 | 17.55 | | | 180.5 | 14.70 | | |
| 9 | Old | 183.56 | 20.26 | 0.963 | 0.0644 | 158.38 | 17.02 | 0.955 | 0.0840 |
| | New | 199.75 | 23.77 | | | 168.06 | 12.18 | | |
| 10 | Old | 139.19 | 13.63 | 0.969 | 0.1897 | 121.06 | 12.60 | 0.953 | 0.0880 |
| | New | 146.69 | 17.81 | | | 129.75 | 14.66 | | |
| 11 | Old | 83.06 | 11.86 | 0.868 | 0.0034 | 75.75 | 11.57 | 0.885 | 0.0081 |
| | New | 104.13 | 14.20 | | | 91.44 | 12.29 | | |
| 12 | Old | 55.31 | 9.13 | 0.868 | 0.0038 | 50.813 | 12.98 | 0.838 | 0.0039 |
| | New | 71.0 | 11.57 | | | 64.19 | 14.06 | | |
| 13 | Old | 28.62 | 5.69 | 0.747 | 0.0033 | 27.19 | 4.97 | 0.745 | 0.0018 |
| | New | 47.44 | 9.08 | | | 40.75 | 7.22 | | |
| 14 | Old | 15.94 | 3.82 | 0.594 | 0.0073 | 15.31 | 3.36 | 0.560 | 0.0074 |
| | New | 25.94 | 6.04 | | | 24.25 | 5.21 | | |
| 15 | Old | 6.20 | 3.96 | 0.668 | 0.0015 | 7.94 | 3.40 | 0.628 | 0.0012 |
| | New | 13.35 | 3.57 | | | 14.56 | 3.77 | | |
| 16 | Old | 3.81 | 2.29 | 0.568 | 0.0037 | 3.75 | 2.24 | 0.557 | 0.0033 |
| | New | 7.69 | 3.20 | | | 7.44 | 2.92 | | |
| 17 | Old | 2.00 | 1.79 | 0.490 | 0.0065 | 2 | 1.79 | 0.549 | 0.0102 |
| | New | 4.38 | 1.86 | | | 4.125 | 1.96 | | |
| 18 | Old | 0.81 | 1.11 | 0.504 | 0.0282 | 0.75 | 1.06 | 0.476 | 0.0278 |
| | New | 1.94 | 1.44 | | | 1.81 | 1.33 | | |
| 19 | Old | 0.50 | 0.89 | 0.347 | 0.0138 | 0.50 | 0.89 | 0.347 | 0.0138 |
| | New | 1.56 | 1.09 | | | 1.56 | 1.09 | | |

*Appendix II.* Kemal logistic regression results comparing token and type lengths between old and new works

| | | Tokens | | | | Types | | | |
|---|---|---|---|---|---|---|---|---|---|
| Word len. | Work type | Average number of of occur. | Stand. dev. | Odds ratio | Prob > Chi-square | Average number of occur. | Stand. dev. | Odds ratio | Prob > Chi-square |
| 2 | Old | 177.63 | 21.66 | 0.971 | 0.0738 | 24.62 | 2.89 | 1.481 | 0.0225 |
| | New | 194 | 26.45 | | | 22.06 | 2.35 | | |
| 3 | Old | 302.63 | 32.82 | 1.020 | 0.1133 | 72.75 | 8.30 | 1.118 | 0.0547 |
| | New | 284.44 | 29.08 | | | 67.25 | 6.09 | | |
| 4 | Old | 303.5 | 25.48 | 1.061 | 0.0096 | 111 | 11.47 | 1.037 | 0.3185 |
| | New | 267.25 | 27.76 | | | 107.31 | 9.31 | | |
| 5 | Old | 459.88 | 45.01 | 1.004 | 0.6668 | 220.94 | 16.60 | 1.021 | 0.4033 |
| | New | 454.19 | 30.31 | | | 216.44 | 13.85 | | |
| 6 | Old | 322.38 | 28.07 | 1.005 | 0.6499 | 194.44 | 17.84 | 0.994 | 0.8120 |
| | New | 317.56 | 33.33 | | | 195.69 | 12.30 | | |
| 7 | Old | 294 | 25.96 | 0.980 | 0.2333 | 212.25 | 20.13 | 0.980 | 0.3494 |
| | New | 304.06 | 20.49 | | | 218.13 | 14.77 | | |
| 8 | Old | 232.81 | 25.92 | 0.999 | 0.9482 | 177.62 | 23.50 | 0.988 | 0.6114 |
| | New | 233.38 | 24.62 | | | 180.25 | 22.87 | | |
| 9 | Old | 162.44 | 19.61 | 0.975 | 0.1768 | 134.13 | 23.16 | 0.971 | 0.2286 |
| | New | 172.31 | 20.92 | | | 140.69 | 14.07 | | |
| 10 | Old | 110.94 | 15.01 | 0.956 | 0.0813 | 92.94 | 16.00 | 0.943 | 0.0709 |
| | New | 122.56 | 18.92 | | | 102.5 | 15.26 | | |
| 11 | Old | 62.25 | 10.36 | 0.941 | 0.1182 | 54.12 | 9.44 | 0.922 | 0.0975 |
| | New | 68.81 | 11.70 | | | 60.38 | 9.65 | | |
| 12 | Old | 38 | 7.76 | 0.898 | 0.0726 | 34.94 | 7.70 | 0.909 | 0.0882 |
| | New | 44.06 | 8.95 | | | 40.31 | 8.39 | | |
| 13 | Old | 17.25 | 5.23 | 0.979 | 0.7436 | 16.19 | 5.08 | 0.979 | 0.7736 |
| | New | 17.88 | 5.88 | | 16.69 | 5.06 | | | |
| 14 | Old | 9.38 | 3.69 | 0.871 | 0.2331 | 9.12 | 3.38 | 0.860 | 0.2140 |
| | New | 10.75 | 2.72 | | | 10.5 | 2.78 | | |
| 15 | Old | 3.81 | 2.40 | 0.810 | 0.1965 | 3.69 | 2.33 | 0.883 | 0.4608 |
| | New | 4.88 | 2.19 | | | 4.25 | 2.05 | | |
| 16 | Old | 2.12 | 1.82 | 1.089 | 0.7222 | 1.94 | 1.53 | 1.133 | 0.6663 |
| | New | 1.94 | 1.18 | | | 1.75 | 0.930 | | |
| 17 | Old | 1.12 | 0.992 | 0.572 | 0.1717 | 0.625 | 0.885 | 0.572 | 0.1717 |
| | New | 0.625 | 0.885 | | | 1.12 | 1.09 | | |
| 18 | Old | 0.312 | 0.602 | 0.515 | 0.2364 | 0.312 | 0.602 | 0.515 | 0.2364 |
| | New | 0.688 | 1.01 | | | 0.688 | 1.01 | | |
| 19 | Old | 0.062 | 0.25 | >999 | 0.9781 | 0.062 | 0.25 | >999 | 0.9781 |
| | New | 0 | 0 | | | 0 | 0 | | |

*Appendix III.* Altan and Kemal logistic regression results comparing most frequent word counts between old and new works

| | | Altan case | | | | Kemal case | | | |
|---|---|---|---|---|---|---|---|---|---|
| Word len. | Work type | Average number of of occur. | Stand. dev. | Odds ratio | Prob > Chi-square | Average number of occur. | Stand. dev. | Odds ratio | Prob > Chi-square |
| ben | Old | N/A* | N/A | N/A | N/A | 11.19 | 5.01 | 1.121 | 0.1009 |
| | New | | | | | 7.81 | 6.03 | | |
| bir | Old | 107 | 15.61 | 1.056 | 0.0818 | 85.81 | 13.10 | 0.976 | 0.399 |
| | New | 98.31 | 9.90 | | | 89.69 | 13.21 | | |
| bu | Old | 26 | 10.84 | 1.168 | 0.0089 | 30.06 | 13.48 | 0.971 | 0.3326 |
| | New | 15 | 5.56 | | | 34.25 | 10.93 | | |
| çok | Old | 7.62 | 4.86 | 0.738 | 0.0242 | 11.12 | 4.75 | 0.897 | 0.1378 |
| | New | 11.12 | 1.82 | | | 14 | 5.75 | | |
| da | Old | 16.75 | 6.88 | 0.668 | 0.0093 | 29 | 9.47 | 0.940 | 0.1748 |
| | New | 34.06 | 6.81 | | | 33.31 | 7.77 | | |
| daha | Old | 13.12 | 5.24 | 0.961 | 0.6445 | 9.88 | 4.72 | 0.819 | 0.0236 |
| | New | 13.81 | 3.15 | | | 14.19 | 4.65 | | |
| de | Old | 23.75 | 7.33 | 0.655 | 0.0091 | 32.19 | 8.46 | 0.910 | 0.0504 |
| | New | 41.38 | 6.30 | | | 38.94 | 8.98 | | |
| dedi | Old | N/A | N/A | N/A | N/A | 21.44 | 10.39 | 1.283 | 0.0026 |
| | New | | | | | 6.88 | 5.56 | | |
| diye | Old | N/A | N/A | N/A | N/A | 9.19 | 3.37 | 1.012 | 0.8936 |
| | New | | | | | 9 | 4.71 | | |
| gibi | Old | 15.75 | 5.01 | 1.372 | 0.0099 | 18.5 | 5.70 | 1.369 | 0.0084 |
| | New | 10.81 | 2.83 | | | 12.69 | 3.01 | | |
| hiç | Old | 6.75 | 2.84 | 1.133 | 0.3324 | N/A | N/A | N/A | N/A |
| | New | 5.75 | 3 | | | | | | |
| her | Old | N/A | N/A | N/A | N/A | 8.25 | 4.24 | 1.008 | 0.9307 |
| | New | | | | | 8.13 | 4.16 | | |
| için | Old | 14.06 | 5.47 | 0.974 | 0.7466 | N/A | N/A | N/A | N/A |
| | New | 14.56 | 3.27 | | | | | | |
| kadar | Old | 8.56 | 2.56 | 1.054 | 0.5678 | 11.25 | 3.64 | 0.651 | 0.0044 |
| | New | 7.75 | 3.43 | | | 19.50 | 5.65 | | |
| ki | Old | 7.88 | 3.47 | 0.907 | 0.2622 | 9.75 | 5.37 | 0.962 | 0.5937 |
| | New | 9.62 | 2.97 | | | 10.69 | 4.79 | | |
| ne | Old | 17.69 | 3.51 | 0.988 | 0.8466 | 17.25 | 5.59 | 0.895 | 0.0798 |
| | New | 18.06 | 4.75 | | | 21.50 | 7.04 | | |
| sonra | Old | 9.62 | 3.01 | 1.226 | 0.0530 | 11 | 3.46 | 1.083 | 0.5371 |
| | New | 6.81 | 2.55 | | | 10.38 | 2.22 | | |
| ve | Old | 40.94 | 7.66 | 1.046 | 0.2425 | N/A | N/A | N/A | N/A |
| | New | 36.94 | 6.54 | | | | | | |

*N/A: Not applicable.

## Note

[1] The most common form of this example used by school children is

"ÇEKOSLOVAKYALILAŞTIRAMADIKLARIMIZDAN MISINIZ?"

Which means, "Are you one of those whom we were unable to convert to a Czechoslovakian?" However, if we choose this example, according to Turkish spelling rules, we have to write the question suffix "misiniz" ("are you") separately and according to our definition of word in this paper (see Section 4) we would have two words.

## References


Altan Ç. (1997–2001) Şeytanın Gör Dediği. Sabah Newspaper (http://www.sabah.com.tr).

Baayen, R.H. (2001) *Word Frequency Distributions*. Kluwer Academic, Dordrecht, Boston.

Baayen H., Van Halteren H., Tweedie F. (1996) Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3), pp. 121–131.

Binongo J.N.G., Smith M.W.A. (1999) The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing*, 14(4), pp. 445–466.

Çiftlikçi, R. (1997) *Yaşar Kemal Yazar-Eser-Üslup*. Türk Tarih Kurumu Basımevi, Ankara.

Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11), pp. 27–34.

Foltz P.W., Dumais S.T. (1992) Personalized Information Delivery an Analysis of Information Filtering Methods. *Communications of the ACM*, 35(12), pp. 51–60.

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

Hakkani-Tür Z. (2000) *Statistical Modeling of Agglutinative Languages*. Ph.D. Dissertation, Department of Computer Engineering, Bilkent University, Ankara, Turkey.

Heaps H.S. (1978) *Information Retrieval Computational and Theoretical Aspects*. Academic Press Inc., New York, New York.

Holmes D.I., Singh S. (1996) A Stylometric Analysis of Conversational Speech of Aphasic Patients. *Literary and Linguistic Computing*, 11(3), 133–140.

Holmes D.I., Forsyth R.S. (1995) The *Federalist* Revisted: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2), pp. 111–1127.

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

Holmes D.I. (1985) The Analysis of Literary Style – A Review. *Journal of the Royal Statistical Society, Series A*, 148(4), pp. 328–341.

Jain A.K., Murty M.N., Flynn P.J. (1999) Data Clustering: A Review. *ACM Computing Surveys*, 30(3), pp. 264–323.

Kemal Y. (1998) *Fırat Suyu Kan Akıyor Baksana*. Adam Yayınları, İstanbul.

Kemal Y. (1971) *Bin Boğalar Efsanesi*. Cem Yayınevi, İstanbul.

Kessler B., Geoffrey N, Schutze H. (1997) Automatic Detection of Text Genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, pp. 32–38.

Köksal A. (1973) *Automatic Morphological Analysis of Turkish*. Ph.D. Dissertation, Hacettepe University, Ankara, Turkey.

Kucera H., Francis W.N. (1967) *Computational Analysis of Present-Day American English*. Brown University Press, Rhode Island.

Martindale C., Tuffin P. (1996) If Homer is the Poet of the *Iliad*, then He may not be the Poet of the *Odyssey*. *Literary and Linguistic Computing*, 11(3), pp. 109–120.

Naci F. (1999) *Yüzyılın 100 Türk Romanı*. Adam Yayınları, İstanbul.

Ney H., Essen U., Kneser R. (1994) On Structuring Probabilistic Dependences in Stochastic Language Modeling. *Computer Speech and Language*, 8(1), pp. 1–38.

Oflazer K. (1994) Two-level Description of Turkish Morphology. *Literary and Linguistic Computing*, 9(4), pp. 137–149.

Redhouse (1979) *Redhouse Yeni Türkçe-İngilizce Sözlük (New Redhouse Turkish-English Dictionary)*. Redhouse Press, İstanbul.

Rudman, J. (1998) The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, 31(4), pp. 351–365.

Sedelow S.Y. (1970) The Computer in the Humanities and Fine Arts. *ACM Computing Surveys*, 2(2), pp. 89–110.

Solak A., Oflazer K. (1993) Design and Implementation of a Spelling Checker for Turkish. *Literary and Linguistic Computing*, 8(3), pp. 113–130.

Stamatatos E., Fakotakis N., Kokkinakis G. (2001) Computer-based Authorship Attribution without Lexical Measures. *Computers and the Humanities*, 35(2), pp. 193–214.

Tallentire D.R. (1972) *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Authorship Attribution*. Ph.D. Thesis, University of Cambridge.

Tanpınar A.H. (1982) *Huzur*. Dergah Yayınları, İstanbul.

TNP (2000) Articles from Turkish Newspapers, www.nlp.cs.bilkent.edu.tr.

Tweedie F.J., Singh S., Holmes D.I. (1996) Neural Network Applications in Stylometry: The Federalist Paper. *Computers and the Humanities*, 30(1), pp. 1–10.

Whissell C.M. (1994) A Computer-program for the Objective Analysis of Style and Emotional Connotations of Prose – Hemingway, Galsworthy, and Faulkner Compared. *Perceptual and Motor Skills*, 79(22), pp. 815–824.

Williams D.C. (1975) Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62(1), pp. 207–212.

Zipf G.K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.