A Detailed Stylometric Investigation of the *İnce Memed* Tetralogy*

JON M. PATTON[1], FAZLI CAN[2]
[1] *Miami Computing and Information Services*
[2] *Computer Science and Systems Analysis Department*
Miami University, Oxford, OH 45056
(pattonjm, canf)@muohio.edu; December 25, 2004
Technical Report #MiamiU-CSA-04-001

Abstract
We analyze four *İnce Memed* novels of Yaşar Kemal using six style markers: "most frequent words," "syllable counts," "word type -or part of speech- information," "sentence length in terms of words," "word length in text," and "word length in vocabulary." For analysis we divide each novel into five thousand word text blocks and count the frequencies of each style marker in these blocks. The principal component analysis results show clear separation between the first two and the last two volumes; the blocks of the first two novels are also distinguishable from each other. The blocks of the last two volumes are intermixed. This parallels the fact that the author planned the last two volumes as three separate novels, but later condensed them into two. The style markers showing the best separation are "most frequent words" and "sentence length". We use stepwise discriminant analysis to determine the best discriminators of each style marker and then use them in cross validation. The related results concur with the principal component analysis results. For example, the cross validation results obtained by "most frequent words" and "sentence length," respectively, provide 87% and 81% correct classification of the text blocks to their corresponding volumes. Further investigation based on multiple analysis of variance (MANOVA) reveals how the attributes of each style marker group distinguish among the volumes.

Keywords: *agglutinative languages, morphological analysis, statistical analysis, stylometry, Turkish literature.*

1. Introduction

Data mining (Fayyad, et al., 1996) for finding hidden characteristics of literary works, or stylometric analysis (Holmes,1985), uses statistical methods based on measurable text attributes that are referred to as style markers (Forsyth, Holmes; 1996). Such studies aim to discover patterns that are usually unconsciously used by the author of a given literary work. The discovered patterns can be used for various purposes, such as author attribution, distinguishing works from each other, or finding the creation time of works. The patterns obtained by stylometric studies may be hard or impossible to acquire by human-based intuitive methods; however, experiments show that objective measures based on style markers can match the literary critical remarks (Whissel, 1994). Similar methods are also used in different fields that involve

* A shorter version of this study is also available as a journal paper: Patton, J. M., Can, F. (2004) "A stylometric analysis of Yasar Kemal's *Ince Memed* tetralogy." *Computers and the Humanities*, 38(4), 457-467.

other kinds of human artifacts, such as architecture, music, painting, software, etc. (Sedelow, 1970; Oman, Cook, 1989).

In this study we provide a stylometric analysis of the *İnce Memed* novels of Yaşar Kemal. (In the rest of the article, although "Kemal" is not the real last name of the author, we will follow the formal writing tradition and will refer to the author with his pen-last name.) Kemal is arguably the most important and well-known writer of the contemporary Turkish literature in international circles; some literary critics regard him as the greatest living Turkish author (Hebért, Tharaud, 1999, p. iix). For example, he was nominated for the Nobel prize in literature twice in 1987 and 2002. Peter Ustinov filmed the first *İnce Memed* novel, with the name "*Memed, My Hawk*," in 1984 and also played one of the major characters (Abdi Ağa). To many Kemal is the most prominent novelist of Turkey. The well-known Turkish literary critic Fethi Naci, in his book on one-hundred noteworthy Turkish novels of the twentieth century reviewed five works of Kemal (one of them was the four volumes of *İnce Memed*) and all together sixty-two different writers (Naci, 1999). The work of Ramazan Çiftlikçi, which is based on his Ph.D. dissertation, is possibly the most detailed study on Kemal's work in Turkish and provides a showcase of the recognition of his works in international circles (Çiftlikçi, 1997, p. 99), and it comes with a comprehensive reference list about his work. Kemal's conversations with Alain Bosquet (Kemal, 1993), and an edited version of these conversations in English (Hebért, Tharaud, 1999) provide author's own views of his life and work; it is an invaluable resource for people working on Kemal. The "Special Issue on Yaşar Kemal" of *Edebiyât* (a journal of middle eastern literature), which was edited by Ahmet Ö. Evin, also provides a pooled useful resource on his writings (*Edebiyât*, 1980).

Kemal was born in 1923 (Kemal, 1993, p.32) in a small village called Hemite (current name is Gökçeli, or Göğçeli) located in the Çukurova region of southern Turkey, which is the setting of his many works. His real name is Kemal Sadık Gökçeli. He developed his writing language in his early ages as he was improvising songs according to the Anatolian tradition of folk bards ("âşık" in Turkish) (Hebért, Tharaud, 1999, pp. 68-69; Kemal 1993, p. 107), and old Turkish folk poets such as Karacaoğlan (Hebért, Tharaud, 1999, pp. 60; Kemal, 1993, p. 99). In his works Kemal usually mixes myth and reality using a unique self-invented poetic language rich in vocabulary (Başgöz, 1980; Oğuzertem, 2003). He has written more than forty volumes including novels, short stories, folklore studies, essays, and journalistic works, etc..

Kemal published his most commonly known work, *İnce Memed* novels of four volumes, between the years of 1955 and 1987 within a thirty-two year period. Actually it took him thirty-nine years (from 1947 to 1986) to complete the work (Çiftlikçi, 1997, p. 143). In some cases their completion involved short intense writing periods. For the first volume the author says that even before writing it he knew the novel by heart, since he thought about it for years (Kemal, 1993, p. 71). He was so unsure about the work that he even did not want to put his name on it when it was first serialized in the *Cumhuriyet* newspaper in 1953-1954 (Kemal, p. 75). However, later it became his most well-known and probably the most commonly read work. It was on the best-sellers list in England (Hebért, Tharaud, 1999, p. 79; Kemal, 1993, p. 113) and has so far been translated into at least thirty languages (Çiftlikçi, 1997, p. 99).

In this study we analyze Kemal's *İnce Memed* tetralogy by using six style markers: 1) sentence length in terms of number of words, 2) the most frequent words, 3) syllable counts in words, 4) word type information (also known as Part of Speech -POS-) based on a statistical method that exploits a morphological analyzer, 5) word length information in the text, and 6) in lexicon. Unlike our previous study on stylometry reported in (Can, Patton, 2004), for the first time we are using the style markers sentence length, syllable counts, and word types. In this exploration our purpose is to check if Kemal has changed his writing style in this tetralogy when objective style markers are used in measuring the style and if so which style marker is the most successful in distinguishing the volumes from each other. We are confident that our results will help other researchers working on stylometry problems in Turkish and in other agglutinative languages.

The rest of the paper is organized as follows. In Section 2 we give a short review of some related works. In Section 3 first we briefly review the morphological structure of Turkish. A brief overview of the novels, description of the test data and experimental design of the study are 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

In stylometry studies writing styles of authors are analyzed using objective measures. For this purpose about 1,000 style markers have been identified (Rudman, 1997). The occurrence patterns of the selected style markers in the text of interest are examined using statistical methods. These patterns are used to resolve stylometric problems, such as authorship attribution, style change, and stylochronometry (i.e., assigning date to work).

In our previous work we studied the writing style change of Kemal and another Turkish author, Çetin Altan, in their old and new works using respectively their novels and newspaper columns (Can, Patton, 2004) using 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 (Binongo, Smith, 1999) are used to 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 on cross validation. The results imply higher time gap may have positive impact in separation and categorization. Our previous work provides the foundation of the current study and here we use three additional style markers: syllable counts, word types, and sentence length. The idea of using the sentence length as a style marker was introduced by Yule (1938).

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 is provided by 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 by Rudman (1997).

An extensively used style marker is the frequency count of "context free" words (or similarly "most frequent words,"and "function words"). For example, Forsyth and Holmes (1996) study 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. The work by Baayen and his co-workers (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.

The text categorization methods as illustrated by Sebastiani (2002) try to assign texts into predefined categories such as known authors. Their aim is automated categorization of texts into predefined categories as we do in this work by using discriminant analysis-based stepwise cross validation. The work of Cambazoğlu (2001) reports text categorization results using written and

spoken Turkish text with stemmed words and various categorization methods (such as k-NN and naïve Bayesian). His study is interesting, since it provides results based on Turkish text. Another stylometric work based on Turkish text Tür (2000) studies authorship attribution using the unigram language model (Ney, et al., 1994). Within the context of his study Tür shows that stemming decreases the categorization accuracy, since it eliminates important stylistic information.

3. Turkish Language Morphology

Turkish belongs to the Altaic branch of the Ural-Altaic family of languages. The Turkish language alphabet 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 a circumflex, the character ^, on top of them. In modern spelling this approach is rarely used.

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

Turkish is an agglutinative language such as 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 derivational suffixes (Lewis, 1967). In agglutinative languages it is possible to have words that would be translated into a complete sentence in languages such as English. For example, "dayanıştırılamayabilecek miymişiz?" would be translated as "is it said that we may not be able to be made to practice mutual aid?" (Lewis, 1967, p. 153).

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-Tur, 2000, p.31).

The study of Turkish morphology as a computation problem can be found in (Köksal, 1973; 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; Hakkani-Tür, et al. 2002).

4. Experimental Environment and Design

4.1. Experimental Environment

4.1.1. *İnce Memed* Tetralogy

İnce Memed is a poor young villager fighting against cruel landlords (or "ağa"s in Turkish). The events of the four volumes take place between 1924 and 1938 (Çiftlikçi, 1997; p. 200). The stories are set in southern Turkey, in the Çukurova plain, foothills and mountains of Taurus range, and partly the Mediterranean cost. Most of the time the story-teller is the narrator (i.e., third person singular); another method which is particularly used in the first volume is making people speak all together as seen in the old Greek tragedies.

The characters and plots of the novels show parallelism. As parallel character examples a few to mention are Hatçe-Seyran, Abdi-Hamza-Mahmut Ağa, Hürü Ana–Kamer Ana, etc. (Çiftlikçi, 1997; p. 205). The killing of Abdi Ağa, at the end of the first volume, and Ali Safa Bey, at the end of the second volume, are written with identical vocabulary and sentence structures (Başgöz, 1980, p. 46). Each novel begins with a description of Çukurova environs; towards the end of each novel Memed kills one or more cruel landlords and disappears. These killing episodes happen twice in the last volume and this can be attributed to the fact that the last two volumes are actually designed as three novels, but later combined into two volumes.

Kıvrak Ali destanında Bayramoğlunun çocukluğunu, babasını, babasının askerden dönmeyişini, kız kardeşinin, anasının başına gelenleri, dağa çıkışını, dağdaki yiğitliklerini, onun dağda Karayılana, Köroğluna, Baba İshaka, Gençosmana eş olduğunu söylüyor, kahramanlıklarını, ermişliğini, çatal yürekli bir yiğit olduğunu öve öve bitiremiyor, Kurtuluş Savaşından sonra onun köyüne çekilişini, bir ermiş yaşamı sürüşünü dile getiriyor, destan burada söz olarak ses olarak kanatlanıyor, sevgiyle, sıcaklıkla, dostlukla, sevinçle, umutla doluyor taşıyor, İnce Memedi öldürme işine gelince de destanın bu parçası önce dehşet bir hüzünde, umutsuzlukta yürüyor, dinleyenleri Bayramoğluna açındırıyor, herkesi onu anlamaya çağırıyor, sonra da destan birden coşuyor, sesiyle, sözüyle bir öfke çığlığına kesiyor, her şey Bayramoğlunu yeriyor, onun bu durumlara düşecek adam olmadığını, onun ölmesi, kara topraklara düşerek bu alçak durumdan kurtulmasını diliyor, sonra birden de hüzünden, öfkeden, yergiden karalamaya, taşlamaya, güldürüye geçiyor, destan Bayramoğlunun ölüsüne yakılan bir gülünç ağıtla bitiyordu.

Figure 1. Longest sentence of all volumes (135 words, *İnce Memed* Volume 4, pp. 407-408).

In the first volume the author uses mostly short sentences, but in later volumes his sentences become longer, and the longest sentence (135 words) of the four volumes appear in volume 4 and

it is given in Figure 1 (later in Section 5.1. we provide detailed sentence length distribution information in Figure 3.d). According to (Naci, 1999, p. 29, 404, 411) Kemal used the same style in all four volumes of his *İnce Memed* novels. However, there are different views on this issue. For example Konur Ertop, another Turkish literary critic, indicated that as time passed the author changed his story telling style in these novels (Çiftlikçi, 1997, p. 184). According to William C. Hickman (1980) Kemal uses a traditional form in the first volume, and in later volumes attempts to write a more modern novel. The Turkish literature scholar Süha Oğuzertem of Bilkent University also indicated that Kemal changed his style in these novels (Oğuzertem, 1987) – we will have more on this statement later in the conclusion.

4.1.2. *İnce Memed* Data for Experimental Design

In this study an individual text word, *token*, is defined as a continuous string of word characters. 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 "word" characters are the Turkish alphabet letters, and the apostrophe sign. [In Turkish the essential use of the apostrophe sign is to separate a proper noun from its suffix(es) (for example, "Memed'in atı", which means "Memed's horse"). However, Kemal does not follow the traditional Turkish punctuation rules and basically does not use the apostrophe sign in his tetralogy.] The versions of "a" and "i" with a ^ on top of them are regarded as different than "a" and "i." The minimum word length is defined as two (word) characters.

In their classical work, Kucera and Francis (Kucera and Francis, 1967 , pp. 365-366), give the average token and type length as 4.74 and 8.13 characters for 1,014,232 words text of different genres in (American) English. In the Kucera-Francis study their word definition is slightly different than our word definition; for example, according to their definition the character "-" can appear in a word and such an approach is uncommon in Turkish. For comparison purposes some numbers are given in Table II for two other novels ("*Bin Boğalar Efsanesi*," and "*Fırat Suyu Kan Akıyor Baksana*") of Kemal that we used in our previous study (Can, Patton, 2004), and a novel of Ahmet Hamdi Tanpınar, "*Huzur*," which was originally published in 1949. We also give numbers for the old and new newspaper columns of Çetin Altan that we used again in our recent study. The Tanpınar case is given to provide a comparison with a significant Turkish author and this work (*Huzur*) is regarded especially significant by literary critics (Naci, 1999, p. 245). The numbers show that in Turkish both the average token and type lengths are longer than what is

observed in Kucera and Francis. Altan's token and type lengths are longer than those of Kemal and Tanpınar. The row for Kemal's all *İnce Memed* novels combined indicates that the average type length increases as the number of tokens increases. This is expected since we sample more shorter words initially, and as we keep accumulating tokens, the new word types we see tend to be longer (Baayen, 2001).

The average type length and average token length of Turkish newspapers is longer than almost all of the Turkish authors listed in Table II. This may be due to the factual content of this medium; the descriptive nature of such content requires the usage of longer words.

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

Text, Date of Publication –where applicable-	No. of Tokens (N)	No. of Types (V)	Avg. Token Length	Avg. Type Length
Kucera and Francis	1,014,232	50,406	4.74	8.13
Turkish Newspapers, 1997-1998	709,121	89,103	6.52	9.28
Ahmet Hamdi Tanpınar, *Huzur*, 1949	97,748	23,407	6.21	8.54
Çetin Altan, Old Columns, 1960-1969	40,000	14,926	6.25	8.10
Çetin Altan, New Columns, 2000	40,000	14,459	6.52	8.51
Yaşar Kemal, *Bin Boğalar Efsanesi*, 1971	66,969	15,491	5.90	8.04
Yaşar Kemal, *Fırat Suyu Kan Akıyor Baksana*, 1998	73,043	16,450	5.99	8.18
İnce Memed [1], 1955	86,457	17,110	5.80	8.01
İnce Memed 2, 1969	107,348	21,146	5.85	8.24
İnce Memed 3, 1983	156,876	26,805	5.81	8.42
İnce Memed 4*, 1987	164,474	28,350	5.91	8.48
İnce Memed 1-4, 1955-1987	515,155	55,394	5.85	8.82

* A translation from Leonardo da Vinci by Murat Belge, which is given at the beginning of the novel, is excluded.

4.2. Experimental Design

4.2.1. Selection of Block Size and Style Markers

For principal component analysis, discriminant analysis, and other tests we needed observations based on fixed size text blocks. We decided that 5,000 is an appropriate block size to be used (Binongo and Smith, 1996, p. 460; Forsyth and Holmes, 1996, p. 164). In this way we had enough number of blocks to have statistically significant results. At the same time we did not want to have excessive number of blocks by choosing a smaller size. In block generation the text has been divided at every 5,000th word; for a given novel the first 5,000 words constituted the

first block and so on. Accordingly, for the volumes 1 through 4, we respectively obtained 17, 21, 31, and 32 blocks.

As we indicated earlier we used six style markers: 1) most frequent words, 2) sentence length in terms of number of words, 3) number of syllables per word counts, 4) word types, 5) word length in text, and 6) word length in vocabulary. For these we have a short explanation for the second, third, and the last two items; for the other two, namely most frequent words and word types, we provide more in depth consideration in the following two subsections.

For sentence length we counted the number of words in each sentence. As end of sentence indicators, we used the period sign, ellipses, and question and exclamation marks. The sentences crossing the block boundaries are assumed to be the member of the block where the sentence ends. In Turkish the number of syllables in a word is the same as the number of vowels in that word. For example, the first word of the tetralogy "duvarın" contains three syllables: "du," "va," "rın." Therefore, it was easy to obtain the syllable counts per word. For word length information we considered the number of characters of all words and unique words of a block.

We discuss the selection of most frequent words and determination of word types of the word stems in the following two sub sections.

4.2.1.1. Determining Most Frequent Words

For determining the most frequent words we counted the number of occurrences of each distinct word (type) and ranked them according to their frequency of appearance in descending order. This is done individually for each volume and at the same time for all volumes combined together. The list of most frequent 50 words of each volume and all volumes together are given in Appendix I. In determining such words we disregarded the context dependent words, mostly the novels' character names or nick names (such as "Topal" Ali). Such words are shown in gray colored cells in Appendix I. When the words are ranked the ones with the same frequency are given the same rank (such as "iki" and "öyle" of volume 1, they both are assigned rank 23, and in this volume when we consider all words at rank 44 we obtain the first most frequent 50 words). In this study the most frequent word rank information is used for no particular reason; however, it is provided to emphasize the fact that more than one word can have the same occurrence frequency.

After this we decided to use the most frequent first 20 words of all volumes combined. These words in alphabetical order are the following: ben, bir, çok, da, daha, de, dedi, diye, gibi, gün,

her, hiç, kadar, ki, ne, onu, onun, sen, and finally sonra. Most of these words also make the top 20 words of the individual volumes. In our previous study (Can, Patton, 2004) we used 15 most frequent words of Kemal's two novels to study the effect of time on his style. However, in the *İnce Memed* experiments, not reported here, we did not have good results, i.e., the set of the chosen 20 words was not able to distinguish the volumes from each other. In other words, these words are not discriminating enough to distinguish the novels from each other. After this observation we decided to consider the most frequent 50 words of each volume and all volumes combined and take the intersection of these 5 sets. We then used the members of the resultant set in our experiments. The list contains 33 words. These words and their English meanings are provided in Table 4. In the experiments we used these words as our most frequent words.

Table III. Most frequent words used in the analysis, obtained by intersecting the top fifty word sets of the four individual volumes and all volumes together

Word No.	Word	English Meaning of Word	R V1	a i V2	n n V3	k* V4	All
1	adam	man, individual	27	42	30	40	33
2	ben	I	10	15	6	13	12
3	beni	me	42	45	48	46	51
4	bile	even	30	30	24	28	27
5	bir	a, one	1	1	1	1	1
6	böyle	so	19	25	21	26	23
7	bu	this	5	4	4	4	4
8	bütün	all	48	31	26	27	29
9	çok	very	15	7	9	9	9
10	da	too	4	3	3	3	3
11	daha	more	13	11	17	15	16
12	de	too	3	2	2	2	2
13	dedi	said	2	6	10	8	5
14	diye	that	8	10	12	12	10
15	geldi	came	24	21	49	29	30
16	gibi	like	7	6	7	7	7
17	gün	day	17	16	32	20	20
18	her	every	18	18	23	18	18
19	hiç	never, none, nothing	22	14	22	21	19
20	kadar	until	11	9	8	6	8
21	ki	that, which, who	12	12	16	10	13
22	mi	adverb of interrogation	21	19	25	32	24
23	ne	what	6	5	5	5	6
24	onu	her, him, it	26	13	13	14	15
25	onun	hers, his, its	33	17	11	11	14
26	öyle	so	23	34	35	37	32
27	sen	you	16	26	14	17	17
28	sonra	hereafter, later	9	8	15	16	11
29	şu	that	38	27	19	31	28
30	üstüne	over	28	26	28	22	25
31	uzun	long	39	22	46	41	36
32	var	there is, there are	14	20	29	24	21
33	ya	then, so	34	35	18	19	22

* Vi: volume i, $1 \leq i \leq 4$, All: all volumes combined.

4.2.1.2. Determining Word Types (Part of Speech Information)

In this study we also used the number of occurrences of the word stem word types (or part of speech, POS, information). For this purpose we first tried to obtain the correct stem of each word and then use the POS information provided by a morphological analyzer indicated for that stem. Study reported by (Altintas, Can, 2002) describes five different stemming approach based on the data derived from the disambiguated Turkish newspaper corpus (see Table II) available from Bilkent University Computer Engineering Department (CTLSP, 2003). Oflazer's (1994) morphological analyzer is used to disambiguate the corpus. The percentage of occurrence of each POS for the word stems of this disambiguated corpus is shown in Table IV. In this table post positions are words like -e kadar (until), -den beri (since) and duplications are words like "ışıl ışıl" (shining brightly). The table shows that more than half of the words are nouns and almost one fifth are verbs. Average stem length for tokens is observed as 4.58 characters (Altintas, Can, 2002).

Table IV. Frequencies for each part of speech (POS) (adapted from Altintas-Can, 2002)

Part of Speech	% Occurrence
Nouns (Nou*)	54.567
Verbs (Ver)	20.023
Adjectives (Adj)	7.955
Conjunctives (Con)	4.867
Numbers (Num)	1.742
Determiners (Det)	3.316
Adverbs (Adv)	2.850
Post Positions (Pos)	2.246
Pronouns (Pro)	2.089
Questions (Que)	0.266
Interjections (Int)	0.060
Duplications (Dup)	0.017

*Abbreviations given in parentheses are used in the text in later sections.

In this study we used one of the statistical stemming algorithm defined by Altintas and Can (2002) to determine the stem of a word. The algorithm uses the morphological analyzer of Oflazer (1994). The Altintas-Can stemming algorithm involves two steps:

1. First it obtains the possible stems of a word using Oflazer's (1994) morphological analyzer (an output of this program is provided in Figure 2). In the results of the analyzer, the first morpheme is the root of the corresponding analysis. It is followed by POS information, and then other morphemes come to form the analysis.
2. In the second step we compare the lengths of the stems of the possible analyses with the average stem length for tokens (which is obtained as 4.58 by Altintas and Can) and choose the stem with the closest length to the average. When there is more than

one result with the same length, the POS of the stem is considered; and the stems are given precedence according to their frequency of occurrence as indicated in Table IV.

In this study we use the POS information corresponding to the selected stem as the word type and use it in our own analysis. Altintas and Can (2002) showed that this approach determines 81% of the POS information and 88% of the stems correctly. Therefore, in the present work we have some incorrectly determined word types.

duvarın
1. duvar+Noun+A3sg+P2sg+Nom
2. duvar+Noun+A3sg+Pnon+Gen

dibinde
1. dip+Noun+A3sg+P3sg+Loc
2. dip+Noun+A3sg+P2sg+Loc

resmim
1. resim+Noun+A3sg+P1sg+Nom
2. resmi+Adj^DB+Noun+Zero+A3sg+P1sg+Nom

aldılar
1. al+Adj^DB+Verb+Zero+Past+A3pl
2. al+Verb+Pos+Past+A3pl

Figure 2. Example output of morphological analyzer for Turkish using the first four words of *İnce Memed* [1] (Oflazer, 1994).

4.2.2. Motivation for the New Style Markers

In our previous study (Can, Patton, 2004), discriminant analyses yielded good to excellent classification rates for the old and new works of Kemal and Altan using the frequencies of token lengths, type lengths, and most frequent word usage as discriminators. For each set of variables a stepwise discriminant analysis was initially conducted to determine the best discriminators for the old and new works of an author. In Altan's works for example 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. Another discriminant analysis using cross validation was conducted with these token lengths as discriminators. This yielded a 97% correct classification rate with 31 cases out of the 32 (16 old and 16 new) being correctly classified. In addition we got 100% correct classification rates in doing a similar set of analyses on Altan's work using type length frequencies and then using frequencies of most frequent word usage.

A similar set of analyses on Kemal's old and new works yielded classification rates that were good but not quite as high as Altans's. Using type lengths as discriminators, 25 out of 32 (78%)

of the cases were correctly classified. A similar result was obtained using token lengths. The best discriminators for Kemal's work was most frequent word usage. 31 out of 32 (97%) old and new text blocks were classified correctly.

We tried the three additional style markers: syllable counts, word types, and sentence length as potential discriminators between the old and new works of Altan and Kemal. The results of these discriminant analyses are summarized in Table V (with the details of new and old cases). All three of the additional style markers provided excellent discrimination results with sentence length providing only one misclassification. The outstanding results of these preliminary experiments gave us motivation to use them in this study.

Table V Correct classification rates between old and new works of Altan and Kemal using the six style markers

	Altan	Works	Kemal	Works
Work Type Data Type	New	Old	New	Old
Token Length Frequencies	100.00 (16)	93.75* (15)	81.25 (13)	75.00 (12)
Type Length Frequencies	100.00 (16)	100.00 (16)	75.00 (12)	81.25 (13)
Most Frequent Word Use Rate	100.00 (16)	100.00 (16)	93.75 (15)	100.00 (16)
Sentence Length	100.00 (16)	100.00 (16)	100.0 (16)	93.75 (15)
Syllable Counts	93.75 (15)	100.00 (16)	87.50 (14)	87.50 (14)
Word Type Frequencies	100.00 (16)	81.25 (13)	93.75 (15)	87.50 (14)

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

5. Experimental Results and Discussion

We first compared the style markers: token length, type length, syllable counts per word, and sentence length for changes across the four volumes. We next performed a principal component analysis for each of our six style markers. Scatterplots of principal component scores for each text block were created to graphically illustrate the differences between the four novels. 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. Through a series of MANOVA analyses, we compared the means of each discriminator across all four volumes to determine if time trends exist. All of these analyses were conducted using the SAS for Windows software, Version 8.2.

5.1. Comparisons of Style Markers across the Four Volumes

We conducted a multiple analysis of variance (MANOVA) to test whether the group of style markers: token length, type length, syllable counts per word, and sentence length significantly changed across the four volumes (most frequent words and word type were not included, since their values are more categorical than numeric). For the selected style markers the average length for each 5000 word block was selected as the response variable and the volume number was the classification variable. The analysis output reported a Wilks Lambda of .1625 which was extremely significant ($p<.0001$). This indicated that the mean values of these style markers change significantly over the four volumes. Table VI summarizes the means and standard deviations of the four markers for each of the four volumes.

Table VI. Means and standard deviations for selected style markers

	Token Length		Type Length		Syllable Counts		Sentence Length	
Volume	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
1	5.80	.081	7.03	.082	2.50	.031	5.33	.287
2	5.85	.113	7.18	.116	2.52	.046	7.46	.580
3	5.81	.143	7.20	.128	2.51	.056	8.82	1.12
4	5.91	.161	7.25	.111	2.55	.067	10.01	1.01

Note that the average type length listed per volume is smaller than the average listed in Table II, since each block may have word types that are common to word types of other blocks and these usually include short words.

Individual Analysis of Variance (ANOVA) were then conducted for each of the four style markers where the average length of the style marker per block was selected as the response variable and the volume number was the classification variable. For token length an ANOVA yielded $F(3,97)=3.62$ ($p=.016$). Using Tukey's Studentized Range Test (HSD), which controls for the Type I experimentwise error rate, the mean token length was found to be significantly at the .05 level between volumes 1 and 4, and between 3 and 4.

The ANOVA for average type length yielded, as expected, much stronger results with $F(3,97)=14.20$ ($p<.0001$). The HSD test showed significant differences at the .05 level between Volume 1 and the other three volumes.

Average syllable counts as the response variable yielded results very similar to token length. Here $F(3,97)=3.43$ ($p=.02$) and there were significant differences between volumes 1 and 4, and between 3 and 4.

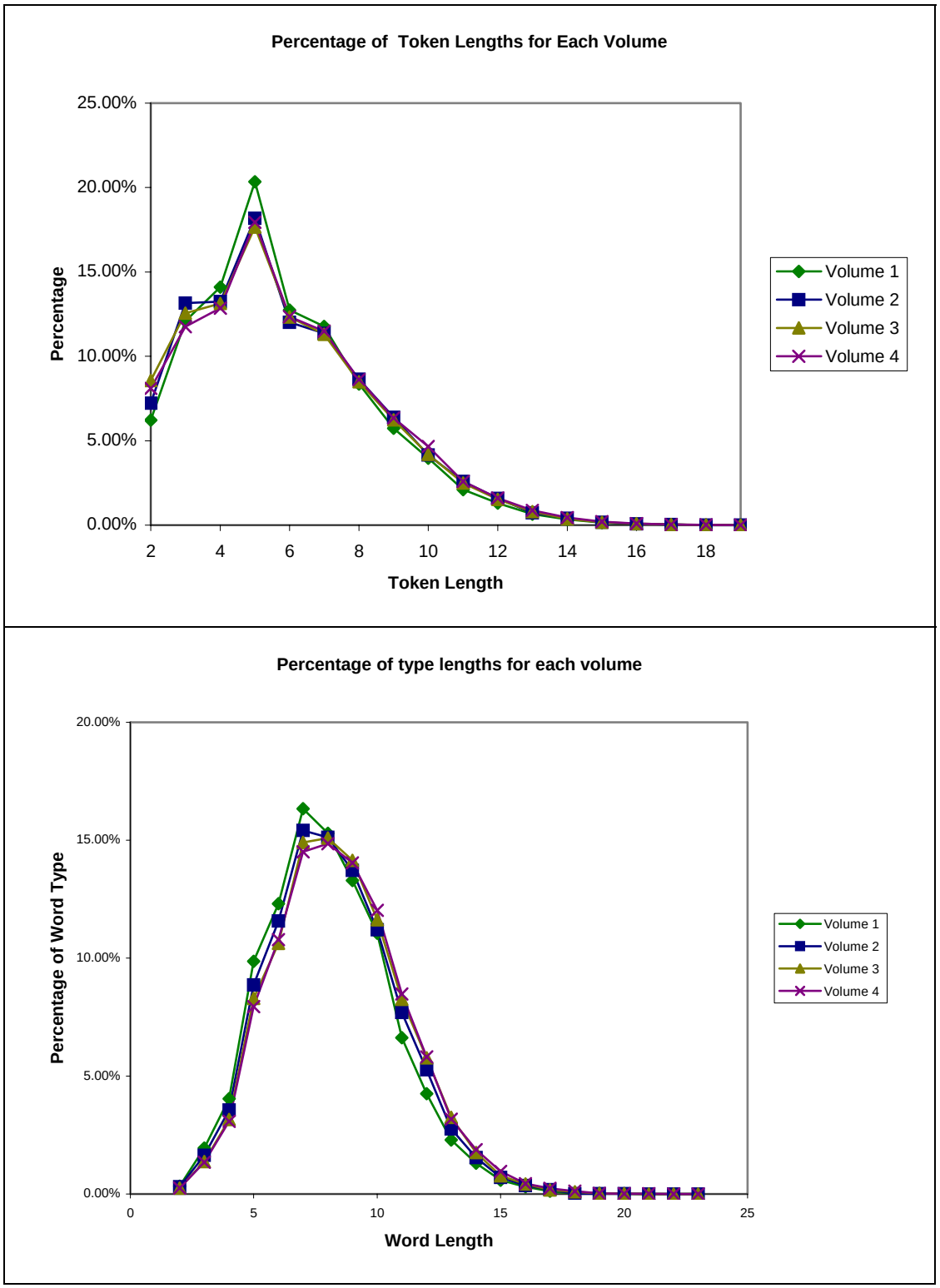

Figure 3.a. Percentage of token lengths for each volume.
Figure 3.b. Percentage of type lengths for each volume.

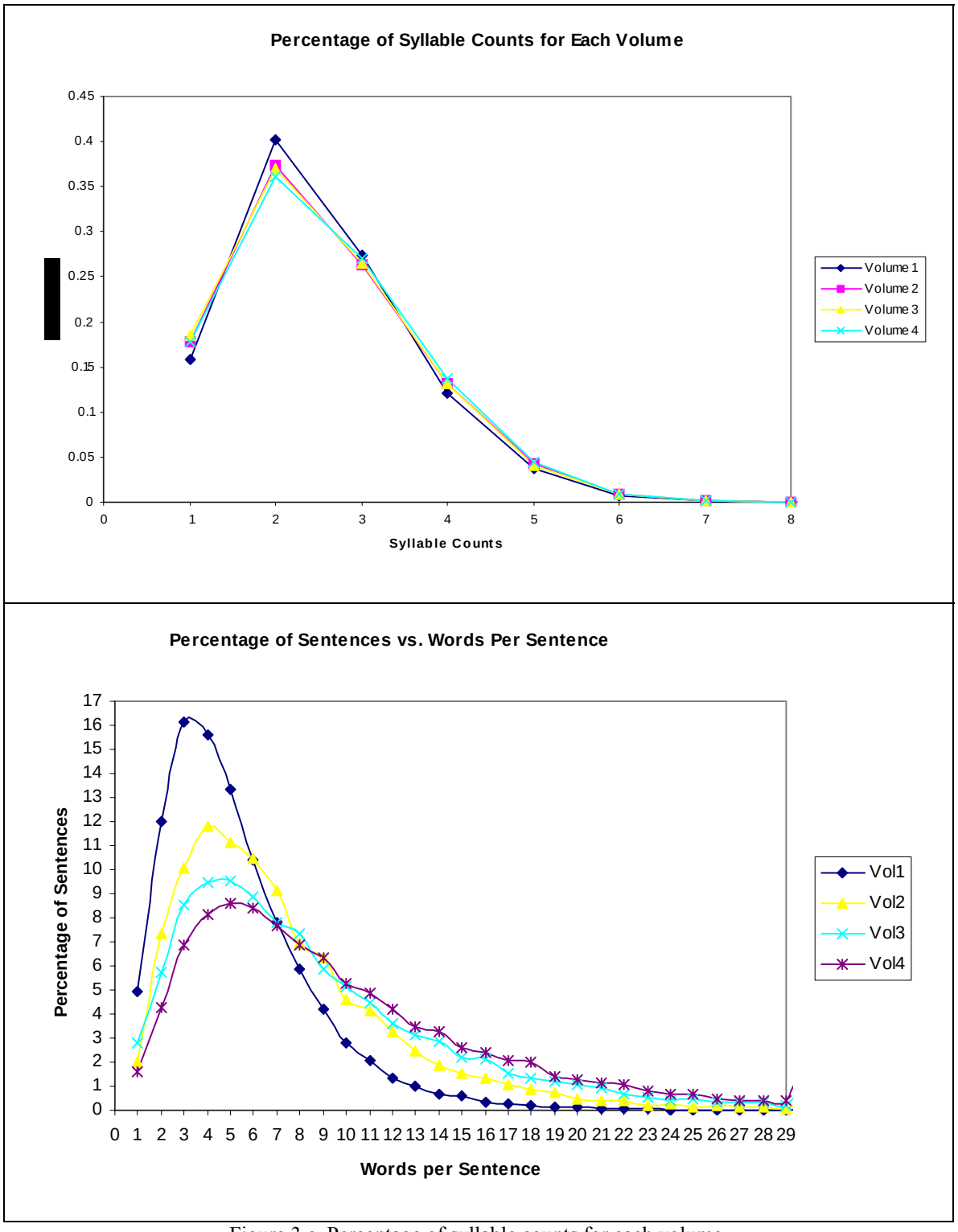

Figure 3.c. Percentage of syllable counts for each volume.
Figure 3.d. Percentage of sentences lengths for each volume.
Figure 3. Percentage distribution information for token lengths, type lengths, syllable counts, and sentence lengths.

The strongest results were generated when average sentence length was the response variable. The ANOVA yielded $F(3,97)=111.20$ ($p<.0001$) and all pair wise differences were significant. Table VI illustrates the significant increase in average sentence length from Volume 1 to Volume 4.

To illustrate comparisons of style marker values for each volume, a series of four plots are presented in Figures 3.a, 3.b, 3.c, 3.d for the style markers "token lengths," "type lengths," "syllable counts," and "sentence lengths." Each plot indicates the percentage usage of each style marker value for each of the four volumes.

Figure 3.b illustrates the relationship between percentages of word types with word length. It is based on all word types per volume instead of word types per block. This is more appropriate, since it is likely that the same word type may appear in several blocks within the same volume. Figure 3.d shows the best separation among the four volumes: Volume 1 has the largest percentage of sentences containing up to five words; Volume 4 has the largest percentage of sentences containing more than 10 words.

5.2. Principal Component Analysis Results

A series of principal component analyses were conducted on the four volumes. The purpose is to transform each of the six sets of related variables (the frequency counts of the most frequent words, sentence length, syllable counts, token lengths, work types, and type length) into a set of uncorrelated variables called principal components (Binongo, Smith, 1999). The plots are presented in Figure 4. All six plots show good separation between the first two volumes, and between volumes 1 and 2 with volumes 3 and 4. The best separation occurred in the plot represented by the frequency counts of the sentence length.

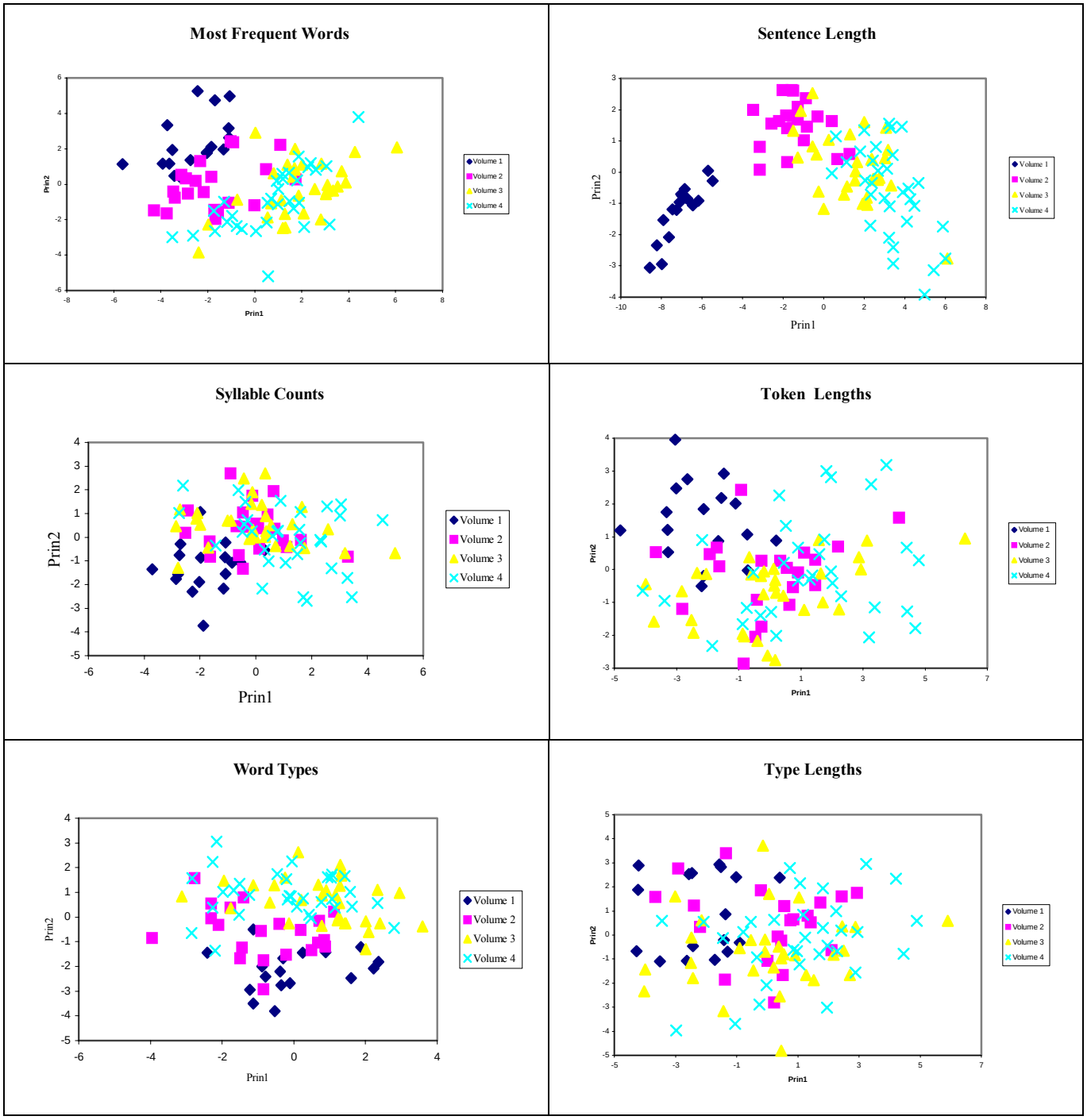

Figure 4. Principal component analyses plots.

5.3. Discriminant Analysis Results

Preceding each discriminant analysis was a stepwise discriminant analysis that determined the best discriminators in each attribute category. The best discriminators among the most frequent words were the following: "dedi", "da", "bir", "çok", "şu", "üstüne", "ya", "var", "bile", "böyle", "sen", "bu", "geldi", "ki", and "onu." Among the syllable counts, the one, two, four, and eight syllable word provided the best separation among volumes. The best discriminators among the word types were con, nou, adj, num, que, pos, det, pro, and ver (Table IV provides the full form of the word type abbreviations).

Using these attribute frequencies as discriminators in each case, 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 in one of the four volumes. Then the excluded block is classified by these linear discriminant functions. This eliminates bias from the classification procedure.

Table VII summarizes the series of discriminant analysis performed on the blocks of text. Each block in the table indicate the percentages of blocks taken from the volume given by the row header (V1, V2, V3, and V4 are respectively volumes 1, 2, 3, and 4) classified as the volume given by the column header. The first row in each block contains the percent classification using discriminators based on sentence length. The second and third row in each block contains the percent classification based respectively on the frequencies of the most frequent words and the syllable counts. The last three rows refer to classification rates for word types, token lengths, and type lengths. For example, the block in the first row and column of the table indicates that, of the 17 blocks of text in Volume 1, all were correctly classified as belonging to this volume based on sentence length. The same is true for the frequencies of the most frequent words. However, 15 out of the 17 (88.24%) were correctly classified using the syllable counts as discriminators, and 76.47% (13 out of 17) of the blocks were correctly classified based on frequencies of each word type. The block of results in the V1 (volume 1) row and V2 (volume 2) column indicates that the 2 blocks of text from volume 1 were incorrectly classified as being in volume 2 based on frequency of syllable counts. The same is true for the 4 misclassified blocks based on frequency of word types.

Table VII. Correct classification rates for each style marker

Novel	Style Marker	V1		V2		V3		V4	
V1 (17)	Sentence Length	100.00%	(17)	0.00%	(0)	0.00%	(0)	0.00%	(0)
	Most Frequent Words	100.00%	(17)	0.00%	(0)	0.00%	(0)	0.00%	(0)
	Syllable Counts	88.24%	(15)	11.76%	(2)	0.00%	(0)	0.00%	(0)
	Word Types	76.47%	(13)	23.53%	(4)	0.00%	(0)	0/00%	(0)
	Token Lengths	94.12%	(16)	5.88%	(1)	0.00%	(0)	0.00%	(0)
	Type Lengths	70.59%	(12)	17.65%	(3)	11.76%	(0)	0.00%	(0)
V2 (21)	Sentence Length	0.00%	(0)	90.48%	(19)	9.52%	(2)	0.00%	(0)
	Most Frequent Words	0%	(0)	90.48%	(19)	0%	(0)	9.52%	(2)
	Syllable Counts	4.76%	(1)	57.14%	(12)	19.05%	(4)	19.05%	(4)
	Word Types	23.81%	(5)	52.38%	(11)	9.52%	(2)	14.29%	(3)
	Token Lengths	14.29%	(3)	80.95%	(17)	4.76%	(1)	0.00%	(0)
	Type Lengths	19.05%	(4)	38.10%	(8)	28.57%	(6)	14.29%	(3)
V3 (31)	Sentence Length	0.00%	(0)	9.68%	(3)	54.84%	(17)	35.48%	(11)
	Most Frequent Words	0.00%	(0)	3.23%	(1)	77.42%	(24)	19.35%	(6)
	Syllable Counts	3.23%	(1)	35.48%	(11)	35.48%	(11)	25.81%	(8)
	Word Types	0.00%	(0)	6.45%	(2)	64.52%	(20)	29.03%	(9)
	Token Lengths	3.23%	(1)	9.68%	(3)	51.61%	(16)	35.48%	(11)
	Type Lengths	12.90%	(4)	25.81%	(8)	32.26%	(10)	29.03%	(9)
V4 (32)	Sentence Length	0.00%	(0)	0.00%	(0)	21.88%	(7)	78.13%	(25)
	Most Frequent Words	0.00%	(0)	6.25%	(2)	12.50%	(4)	81.25%	(26)
	Syllable Counts	0.00%	(0)	18.75%	(6)	21.88%	(7)	59.38%	(19)
	Word Types	0.00%	(0)	12.50%	(4)	28.13%	(9)	59.38%	(19)
	Token Lengths	6.25%	(2)	9.38%	(3)	34.38%	(11)	50.00%	(16)
	Type Lengths	9.38%	(3)	21.88%	(7)	21.88%	(7)	46.88%	(15)

Average for sentence length: 80.86%, most frequent words: 87.29%, syllable counts: 60.06%, word types: 63.19%, token lengths 69.17%, type lengths: 46.95%.

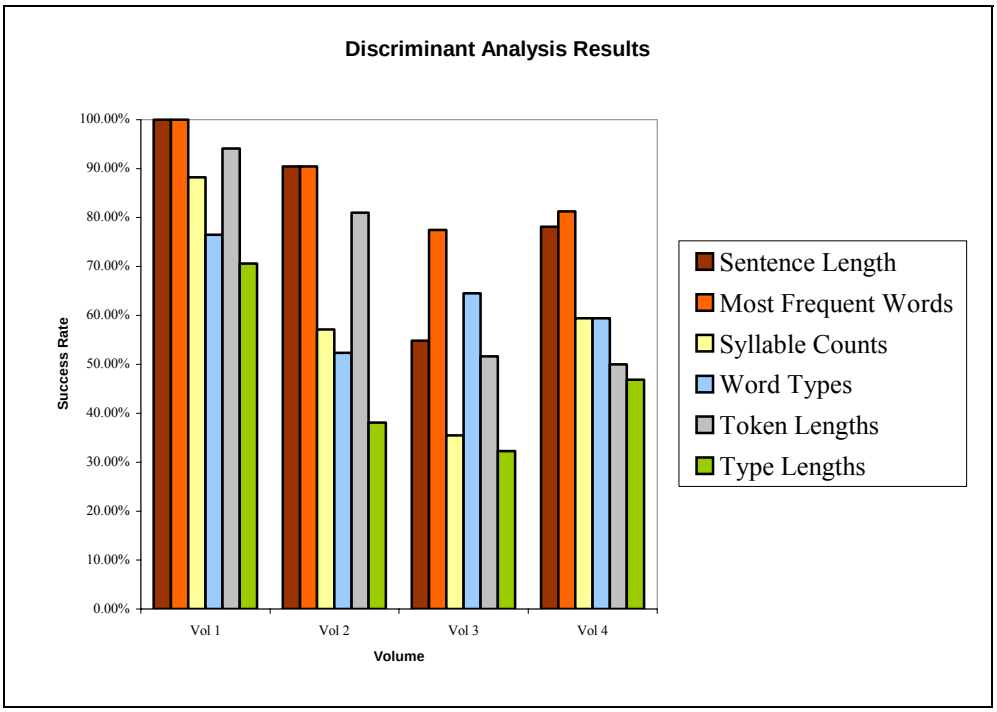

Figure 5. Bar chart of discriminant analysis results of correct classification rates.

Table VII contains the overall correct classification rates for each of the attributes. 87.29% of the blocks were correctly classified using the frequency of the most frequent context free words. Using syllable counts, 60.06% of the blocks were correctly classified. And 63.19% were correctly classified using frequency of word types.

To gain a better perspective of these results, Figure 5 provides a bar chart where each bar type measures the correct classification percentage of a style marker for each volume. It is clear from this chart for example that the style markers sentence length and most frequent words have the highest percentage success of classifying Volume 1 and Volume 2 blocks correctly.

5.4. Manova Results

To determine the volumes that were discriminated by each style marker discriminator, a Multiple Analysis of Variance (MANOVA) was conducted for each style marker using the best discriminators for that marker as the set of dependent variables and the volume as the classification variable. For example the best discriminators of the style marker, sentence length, are sentences of length 1, 2, 3, 5, 6, 7, 9, 11, 12, 16, 22, and 30 words. Denoting SLi (SL1, SL2, etc.) as the number of sentences of length i, Table VIII.a contains the average number of sentences of length i per block for each volume.

Table VIII.a. Means of sentence length per block for each volume

Volume	SL1*	SL2	SL3	SL5	SL6	SL7	SL9	SL11	SL12	SL16	SL22	SL30
1	46.4	112.6	151.7	125.0	97.7	73.4	39.3	19.2	12.6	3.2	.411	.529
2	13.3	49.1	67.7	74.5	70.2	61.1	42.5	27.6	22.0	9.0	2.6	3.48
3	15.9	32.6	48.6	54.4	50.4	44.5	33.4	25.3	20.5	12.1	3.7	7.7
4	7.8	21.0	33.8	42.3	41.4	37.8	31.2	24.0	20.8	11.9	5.2	12.7

- SL1 as the number of sentences of length 1, SL2 represents sentence of length 2 etc.

The two sub-tables making up Table VIII.b contain the average number per text block of the best discriminators among the most frequent words. Again this is done for each volume.

Table VIII.b. Means of frequencies of most frequent words per block for each volume

Volume	ben	beni	bir	böyle	bu	çok	da	de	dedi
1	22.5	8.4	152.2	13.2	45.7	16.9	50.6	53.2	71.6
2	16.9	7.4	184.8	11.4	54.0	29.0	58.7	68.8	29.8
3	30.1	7.8	155.5	14.0	65.5	28.8	64.0	85.5	28.2
4	22.4	74.	144.0	12.3	64.6	27.0	76.5	79.1	28.8

Table VIII.b. (cont) Means of frequencies of most frequent words per block for each volume

Volume	geldi	ki	onu	sen	şu	üstüne	var	ya
1	12.4	17.8	11.4	15.2	8.7	10.8	17.2	9.3
2	11.9	19.6	18.6	11.4	11.1	11.3	13.0	8.8
3	7.6	21.2	26.1	23.2	15.1	12.1	11.8	15.4
4	11.4	25.5	21.8	17.2	10.5	13.8	12.8	15.5

The best discriminators among the syllable counts are provided in Table VIII.c. The values in the SYL1 up through SYL9 column respectively represent the average number of words having 1 up thru 9 syllables per text block for each volume.

*Table VIII.c. Means of syllable counts per block for each volume

Volume	SYL1*	SYL2	SYL3	SYL4	SYL5	SYL6	SYL7	SYL8	SYL9
1	791.1	2013.4	1366.2	602.8	183.9	36.2	5.8	0.47	0.12
2	893.4	1869.8	1315.0	658.4	213.6	42.6	6.6	0.43	0.10
3	918.2	1844.4	1320.5	658.1	209.0	41.2	7.6	0.83	0.13
4	886.3	1802.4	1352.0	680.8	222.8	45.4	8.5	1.41	0.28

Table VIII.d presents the average number per block of the best discriminators among the word types. These abbreviations respectively expand (as we go from left to right) as nouns, verbs, adjectives, conjunctives, numbers, determiners, adverbs, post positions, pronouns, questions, interjections, and duplications. In fact all twelve parts of speech listed in Table IV served as excellent discriminators.

*Table VIII.d. Means of word type frequencies per block for each volume

Volume	Nou	Ver	Adj	Con	Num	Det	Adv	Pos	Pro	Que	Int	Dup
1	2082.2	1334.8	607.8	94.0	66.4	76.3	143.4	77.6	90.5	19.2	10.6	6.9
2	2140.4	1285.4	651.3	108.6	78.2	89.6	136.0	70.6	79.8	17.5	7.1	4.9
3	2102.8	1259.9	648.0	149.5	86.8	104.5	143.0	73.8	104.5	23.7	9.5	3.8
4	2166.8	1255.1	611.9	146.9	86.3	103.3	136.4	69.7	96.7	15.1	7.9	4.7

In table VIII.e, TOKi represent the number of word tokens of length i. The best discriminators are tokens of length 2, 3, 4, 5, 10, 11, 14, 17, and 18. Again the value in each column TOKi represents the average number of word tokens

Table VIII.e. Means of token length per block for each volume

Volume	TOK2	TOK3	TOK4	TOK5	TOK10	TOK11	TOK14	TOK17	TOK18
1	311.1	604.5	705.3	1016.6	198.7	105.0	16.8	1.4	0.12
2	361.2	658.0	662.7	909.2	207.6	129.7	21.0	2.1	0.14
3	428.8	627.3	656.9	883.2	209.5	124.5	18.7	1.8	0.74
4	405.0	587.3	643.0	897.3	232.5	129.3	22.0	2.1	1.03

In Table VIII.f , TYPi represents the number of word types of length i in each block. The best discriminators are types of length 4, 7, 9, 10, and 18.

For each style marker, the Wilks Lambda statistic for the associated MANOVA was significant at a p-value less than .0001. Individual analysis of variance were conducted for each discriminator of that style marker with volume as the classification variable. Table IX summarizes the results of the significant means comparisons for each discriminator using a Tukey's Standardized Range Test. For example the average number of sentences of length 1, 2, 3, 5, 6, 7, 11, 12, 16, and 22 per text block were significantly different ($p<.05$) between Volumes 1 and 2. The sentence length marked in bold text, namely lengths 2, 3, 5, 6, and 7 had average numbers that were significantly different among "all" the volumes.

Table VIII.f. Means of type length per block for each volume

Volume	TYP4	TYP7	TYP9	TYP10	TYP18
1	183.3	375.2	222.6	162.4	0.12
2	178.9	376.9	248.3	172.6	0.10
3	173.7	362.2	245.5	173.3	0.74
4	174.0	373.1	250.3	184.4	1.03

Table IX Means comparisons for each style marker*

Volume Pair	Sentence Length **	Most Frequent Words	Syllable Counts	Word Types	Token Lengths	Type Lengths
V1-V2	SL1,**SL2,SL3 SL5,SL6,SL7** SL11,SL12,SL16 SL22	bir, çok,de dedi,onu, sen	SYL1,SYL2, SYL4,SYL5	Adj	TOK2,TOK3, TOK5,TOK11	TYP9
V1-V3	SL1**,SL2,SL3 SL5,SL6,SL7** SL9,SL11,SL12 SL16,SL22,SL30	bu, çok,da de,dedi, geldi,onu sen,şu,var, ya	SYL1,SYL2, SYL4	Ver,Adj, Con,Num, Dup	TOK2,TOK4, TOK5,TOK11	TYP4,TYP9
V1-V4	SL1,**SL2,SL3 SL5,SL6,SL7** SL9,SL11,SL12 SL16,SL22,SL30	bu, çok,da de,dedi,ki onu,ya	SYL1,SYL2, SYL4,SYL5, SYL6,SYL8	Nou,Ver, Con,Num	TOK2,TOK4, TOK5,TOK10, TOK11,TOK14 TOK18	TYP4,TYP9, TYP10,TYP18
V2-V3	**SL2,SL3,SL5 SL6,SL7**,SL9 SL16,SL30	ben,bir,bu da,de,geldi onu,ya	-	Con,Pro	TOK2	TYP7
V2-V4	**SL2,SL3,SL5 SL6,SL7**,SL9 SL11,SL16,SL22 SL30	bir,bu,da de,ya	SYL2,SYL8	Adj,Con	TOK2,TOK3, TOK10,TOK14	TYP18
V3-V4	SL1,**SL2,SL3 SL5,SL6,SL7** SL22,SL30	ben,geldi, sen, şu	-	Nou,Adj, Que	TOK3,TOK10,	-

* Empty cells are indicated by the - symbol.
** Bold sentence lengths had significant differences among all the volumes.

It is interesting to note that by inspecting Table VIII.d in conjunction with Table IX the average number of nouns per block is significantly higher in volume 4 when compared with volumes 1

and 3. This is in tandem with the increase in the average number of sentences per block of length 22 and 30 between these same volumes. This possibly can be explained by noting that longer sentences are expected to have a larger concentration of nouns than shorter sentences.

6. Conclusions

In this study we analyze the *İnce Memed* tetralogy of Yaşar Kemal using six style markers: "most frequent words," "syllable counts," "word type -or part of speech- information," "sentence length in terms of words," "word length in text," and "word length in vocabulary." Sentence length, syllable counts, and word types are used for the first time to analyze Turkish text. They were selected because in our preliminary experiments they provided excellent discrimination results between the old and new works of both Altan and Kemal as illustrated in Table V.

Principal component and discriminant analyses were conducted to determine the separation between the novels using these six style markers. Although all six style markers provided strong results in the separation and classification of volumes 1 and 2, sentence length and most frequent words provided excellent overall results in distinguishing among all four volumes. The principal components based on sentence length and most frequent words provided excellent separation between the works as illustrated in Figure 4. We used stepwise discriminant analysis to determine the best discriminators of each style marker and then used them in discriminant analysis based on cross validation. The classification rates are presented in Table VII. The successful categorization rate for the six style markers are as follows: 81% for sentence length, 87% for most frequent words, 60% for syllable counts, 63% for word types, 69% for token lengths, and 47% for type lengths. Although all the style markers did not yield the correct classification rate levels as the ones presented in Table V, note that we are using four classification levels in this study instead of the two levels used previously. However, the percentage classification success rate for both sentence length and most frequent words as discriminators are outstanding for these four volumes.

Among the new three style markers, sentence length appears to provide the best discrimination among the four volumes. Our MANOVA results in Table IX provide the sentence lengths that are the best discriminators: the sentence lengths of 2, 3, 5, 6, and 7 words discriminate among all the volumes. Furthermore, Table VI illustrates that the average sentence length per block increases substantially as we go from volume 1 to volume 4. The average (per block) length of word types also increases, but not as dramatically as sentence length. There may be a

correspondence between word types and sentence length, since we observed that a larger number of nouns per sentence were good discriminators between volumes 1 and 4, and 3 and 4. This was also true for larger sentence lengths.

The word type determination algorithm produces correct results 80% of the time. An improved algorithm such as the one referenced in (Hakkani-Tür, et al., 2003), may provide more accurate results. In turn this could provide better classification success rates for word types. These require additional investigation.

The key contribution of this study is the outstanding success levels achieved using various style markers in distinguishing changes of a Turkish author's writing style; furthermore, this article is the first comprehensive stylometric study exclusively devoted to Kemal. The style markers of this study and the insight brought by their success can be valuable information for other researchers; they may exploit these markers in their stylometric investigation of various aspects of the Turkish language and other agglutinative languages.

Previous experiments show that objective measures based on style markers can match the literary critical remarks (Whissel, 1994). Our results show clear separation between the first two and the last two volumes. The blocks of the first two novels are also distinguishable from each other; and the blocks of the last two volumes are intermixed. This parallels the fact that the author planned the last two volumes as three separate novels, but later condensed them into two. More interestingly, Oğuzertem (1987) states that the first novel could be termed "romantic," the second "realistic," and the last two "postmodernist." The nature of his comments is interesting, since our objective results are in harmony with them, although we approach the matter from different perspectives. This separation can also be attributed to the change of style with time (Can, Patton, 2004).

In our future work we are planning to investigate the Turkish literature from a broad perspective by studying various Turkish authors from different periods of the 20th century using the same style markers and assess the markers' strengths in terms of data mining.

Acknowledgements

We greatly appreciate the detailed comments on the article made by Süha Oğuzertem of the Turkish Literature Department, Bilkent University; we also thank him for bringing his 1987 conference paper and presentation to our attention. We are grateful to Engin Demir of Bilkent University for his generous time offer to this project in preparing the input data sets. We thank

Kemal Oflazer of Sabancı University for making the morphological analyzer and the Turkish newspaper collection available. One of the authors is grateful to the Computer Engineering Department of Bilkent University, since his sabbatical leave there greatly enabled the realization of this project.

Appendix I. Most Frequent 50 words of each volume and all volumes combined (*)

Raw Rank	V1	V2	V3	V4	All (V1-V4)
1	1, bir, 2634	1, bir, 3971	1, bir, 4885	1, bir, 4759	1, bir, 16249
2	memed, 1245	2, de, 1472	2, de, 2676	2, de, 2586	2, de, 7646
3	2, dedi, 1231	3, da, 1275	3, da, 2632	3, da, 2509	3, da, 7286
4	3, de, 912	4, bu, 1141	4, bu, 2048	4, bu, 2094	4, bu, 6072
5	4, da, 870	5, ne, 755	ince, 1438	ince, 1341	memed, 4213
6	5, bu, 789	memed, 681	5, ne, 1138	memed, 1323	5, dedi, 3716
7	6, ne, 624	6, **dedi**, 644	memed, 964	5, ne, 1096	6, ne, 3613
8	ali, 583	6, gibi, 644	6, ben, 946	6, kadar, 965	ince, 3449
9	7, gibi, 566	7, çok, 615	7, gibi, 898	7, gibi, 953	7, gibi, 3061
10	8, diye, 502	8, sonra, 592	8, kadar, 897	8, dedi, 952	8, kadar, 2726
11	9, sonra, 488	ali, 542	9, çok, 894	9, çok, 878	9, çok, 2676
12	cabbar, 428	koca, 535	10, dedi, 889	10, ki, 833	10, diye, 2617
13	abdi, 407	bey, 529	11, onun, 842	11, onun, 786	11, sonra, 2440
14	10, ben, 386	9, kadar, 517	12, diye, 825	12, diye, 781	12, ben, 2437
15	11, kadar, 347	10, diye, 509	13, onu, 816	13, ben, 745	ali, 2353
16	çavus, 342	11, daha, 453	ali, 795	14, onu , 713	13, ki, 2227
17	ağa, 310	12, ki, 421	ağa, 743	15, daha, 707	14, onun, 2128
18	12, ki, 306	13, onu, 399	14, sen, 721	16, sonra, 668	15, onu, 2124
19	ince, 305	osman, 390	15, sonra, 692	17, sen, 571	16, daha, 2123
20	13, daha, 302	14, hic, 380	16, ki, 667	18, her, 549	17, sen, 1796
21	14, var, 299	safa, 372	17, daha, 661	bey, 504	18, her, 1511
22	durdu, 299	ince, 365	murtaza, 568	19, ya, 504	19, hiç, 1493
23	15, çok, 289	15, ben, 360	18, ya, 484	20, gün, 471	ağa, 1472
24	topal, 265	16, gün, 349	19, şu, 468	21, hiç, 456	bey, 1409
25	16, sen, 262	17, onun, 338	20, ona, 448	22, üstüne, 453	20, gün, 1409
26	17, gün, 235	18, her, 300	21, böyle, 439	hoca, 452	21, var, 1363
27	18, her, 230	19, mi, 285	22, hiç, 437	23, ona, 449	22, ya, 1333
28	19, böyle, 226	ana, 283	23, her, 432	ali, 433	23, böyle, 1321
29	20, yok, 225	20, var, 278	24, bile, 424	memedin, 433	24, mi, 1270
30	21, mi, 223	at, 277	25, mi, 420	24, var, 419	25, üstüne, 1263
31	22, hiç, 220	21, geldi, 263	26, bütün, 405	25, ve, 411	26, ona, 1261
32	hatçe, 220	22, uzun, 258	27, benim, 395	26, böyle, 406	27, bile, 1210
33	23, **iki**, 217	23, iki, 256	28, üstüne, 381	27, bütün, 401	28, şu, 1202
34	23, öyle, 217	24, ama, 254	29, var, 367	28, bile, 392	29, bütün, 1160
35	24, geldi, 216	25, böyle, 250	30, adam, 362	29, geldi, 383	memedin, 1138
36	memedin, 214	26, **sen**, 242	31, şimdi, 356	30, benim, 349	30, geldi, 1105
37	25, doğru, 202	26, üstüne, 242	32, gün, 354	at, 346	31, benim, 1054
38	26, onu, 196	27, **içinde**, 236	33, ve, 339	31, şu, 346	32, öyle, 1001
39	27, adam, 193	27, şu, 236	topal, 337	32, mi, 342	ana, 1001
40	recep, 187	28, gene, 229	hürü, 331	33, senin, 337	33, adam, 997
41	28, üstüne, 187	29, ona, 228	34, senin, 328	murtaza, 323	at, 973
42	29, **benim**,181	30, bile, 217	ana, 319	34, için, 314	34, şimdi, 958
43	29, işte, 181	kamer, 217	at, 317	ana, 311	35, senin, 931
44	30, bile, 177	31, bütün, 212	yüzbaşı, 303	35, **değil**, 310	36, uzun, 928
45	31, gene, 175	32, şey, 210	35, öyle, 301	35, şimdi, 310	37, içinde, 922

(*) Each cell contains the following information: frequency rank, word, and word occurrence frequency in text (context dependent words are grayed and not considered for ranking). Words with the same frequency are assigned the same rank; the first of such words is shown in bold.

Appendix I (cont.). Most Frequent 50 words of each volume

Raw Rank	V1	V2	V3	V4	All
46	32, iyi, 166	33, gece, 209	36, mı, 300	36, içinde, 301	38, iki, 921
47	33, **onun**, 162	arif, 204	37, işte, 298	37, öyle, 294	murtaza, 921
48	33, sordu, 162	saim, 200	38, için, 297	ağa, 283	topal. 902
49	süleyman, 159	memedin, 194	memedin, 297	arif, 282	39, için, 895
50	34, ya, 159	34, öyle, 191	39, bana, 290	38, hiçbir, 278	40, gene, 892
51	35, mı, 158	35, ya, 186	memedi, 290	39, büyük, 275	41, işte, 883
52	36, içinde, 157	seyran, 185	bey, 288	ferhat, 270	42, değil, 877
53	37, şimdi, 153	36, hiçbir, 184	40, iyi, 285	40, adam, 269	43, ve, 876
54	38, şu , 152	37, vardı, 183	41, **değil**, 277	anacık, 267	44, yok, 853
55	ıraz, 149	38, üç, 182	41, sana, 277	41, uzun, 266	45, iyi, 849
56	39, **başladı**, 148	ferhat, 180	mahmut, 263	atı, 264	46, mı, 843
57	39, vardı, 148	39, güzel, 177	42, gene, 262	saim, 260	47, hiçbir, 827
58	39, uzun, 148	40, büyük, 176	43, hem, 260	42, üç, 257	48, şey, 826
59	40, oldu, 146	41, hem, 175	44, seni, 259	43, işte, 256	49, sana, 813
60	41, şey, 145	42, adam, 173	45, nasıl, 258	bayramoğlu, 255	koca, 806
61	deli, 144	beyin, 171	46, uzun, 256	44, sana, 247	50, ama, 798
62	42, beni, 143	43, doğru, 171	47, hiçbir, 251	45, en, 246	-
63	43, bütün, 142	44, değil, 167	tazı, 251	46, **beni**, 245	-
64	vay, 141	45, beni, 163	48, beni, 244	46, iyi, 245	-
65	44, **ama**, 140	46, yok, 161	-	47, yok, 243	-
66	44, başını, 140	47, **gitti**, 160	-	48, şey, 231	-
67	44, sana, 140	47, nasıl, 160	-	-	-
68	-	48, **hemen**, 159	-	-	-

References

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

Altintas, K., Can, F. (2002) "Stemming for Turkish: A Comparative Evaluation". In the *Proceedings of 11th Turkish Symposium on Artificial Intelligence and Neural Network* (İstanbul, 20-21 June 2002), pp.181-188.

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

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

Başgöz, İ., (1980) "Yaşar Kemal and Turkish Folk Literature." *Edebiyât*, 5(1-2), 37-47.

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.

Cambazoğlu, B. B. (2001) "Automatic Text Categorization on Turkish Text Documents." Last accessed on April 6, 2003, < http://www.cs.bilkent.edu.tr/~berkant/publications.html>.

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

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

CTLSP (Center for Turkish Language and Speech Processing). (2003) Turkish Texts: Articles from Turkish Newspapers. Last accessed on April 5, 2003, <http://www.nlp.cs.bilkent.edu.tr/Center/Corpus/>.

Edebiyât. (1980) Special issue on Yaşar Kemal, edited by A. Ö. Evin, 5(1-2), 238 pages.

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), 27-34.

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

Hakkani-Tür Z., Oflazer, K, Tür, G. (2002) "Statistical Morphological Disambiguation for Agglutinative Languages. *Computers and the Humanities*, 36(4), 381-410.

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

Hebért, E. L., Tharaud, B. (1999) *Yaşar Kemal on His Life and Art*. Syracuse University Press, Syracuse, NY.

Hickman, W. C. (1980) "Traditional themes in the work of Yaşar Kemal: İnce Memed." *Edebiyât*, 5(1-2), 55-68.

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

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

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

Kemal, Y. (1993) *Yaşar Kemal Kendini Anlatıyor*. Toros Yayınları, İstanbul.

Kemal Y. (1987) *İnce Memed 4*. Toros Yayınları, İstanbul.

Kemal Y. (1983) *İnce Memed 3*. Toros Yayınları, İstanbul.

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

Kemal Y. (1969) *İnce Memed 2*. Ant Yayınları, İstanbul.

Kemal Y. (1955) *İnce Memed* [1]. Remzi Kitabevi, İstanbul.

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.

Lewis, G. L. (1967) *Turkish Grammar*. Oxford University Press, Glasgow.

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), 1-38.

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

Oğuzertem, S. (2003) "Geçmisten Geleceğe Yaşar Kemal." Uluslararası Yaşar Kemal Sempozyumu, Açılış Bildirisi, Bilkent Universitesi, Ankara. In Oğuzertem, S., ed. *Geçmisten Geleceğe Yaşar Kemal*. Adam Yayınları, İstanbul, 25-41.

Oğuzertem, S. (1987) "Yashar Kemal's *İnce Memed*'s: Myth in the Making." Indiana University, Bloomington, Indiana.

Oman, P., Cook, C. "Programming style authorship analysis". In *Proc. 17th Annual ACM Computer Science Conference,* pp. 320-326.

Rudman, J. (1997) "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.

Sedelow, S. Y. (1970) "The Computer in the Humanities and Fine Arts." *ACM Computing Survey*s, 2(2), 89-110.

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

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

Tür, G. "Automatic Authorship Detection" Last accessed on May 6, 2003, <http://www.research.att.com/~gtur/pubs/authorship.pdf>.

Whissel 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), 815-824.

Yule, G.U., (1938) "On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship." *Biometrika*, 30, 363-390.