

A Stylometric Analysis of Yaşar Kemal's *İnce Memed* Tetralogy

JON M. PATTON¹, FAZLI CAN²

¹ Information Technology Services

² Computer Science and Systems Analysis Department

Miami University, Oxford, OH 45056

(pattonjm, canf)@muohio.edu; April 5, 2004

Tentative final form submitted to the journal, to appear in "Computers and the Humanities"

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 style markers showing the best separation are "most frequent words" and "sentence lengths." We use stepwise discriminant analysis to determine the best discriminators of each style marker. We then use these markers in cross validation based discriminant analysis. 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.*

1. Introduction

Data mining for finding hidden characteristics of literary works, or stylometric analysis, 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. In this study we provide a stylometric analysis of the *İnce Memed* tetralogy of Yaşar Kemal. He is arguably the most important and well-known writer of the contemporary Turkish literature (Hebért, Tharaud, 1999, p. iix). The work of Çiftlikçi (1997, p. 99) provides a showcase of the recognition of his works in international circles. Kemal published his most commonly known work, *İnce Memed* novels of four volumes, between the years of 1955 and 1987. We analyze the *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 stemmer that exploits a morphological analyzer, 5) word length information in the text, and 6) word length information in the lexicon. In this exploration our purpose is to check if Kemal has changed his writing style in this tetralogy when objective style markers are used; and, if so, which style marker is the most successful in distinguishing the volumes from each other.

2. Experimental Environment and 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 is case insensitive. The “word” characters are the Turkish alphabet letters, and the apostrophe sign (not used by Kemal). The versions of “a” and “i” with a ^ on top of them are regarded as different than “a” and “i” (these versions of “a” and “i” are also not used by Kemal). The minimum word length is defined as two (word) characters. Token and type statistical information for the novels of the tetralogy is provided in Table I.

Table I. No. of tokens, types, and their length information for the *İnce Memed* tetralogy

Novel, Date of Publication	No. of Tokens (N)	No. of Types (V)	Avg. Token Length	Avg. Type Length
<i>İnce Memed</i> [1], 1955	86,457	17,110	5.80	8.01
<i>İnce Memed</i> 2, 1969	107,348	21,146	5.85	8.24
<i>İnce Memed</i> 3, 1983	156,876	26,805	5.81	8.42
<i>İnce Memed</i> 4, 1987 (*)	164,474	28,350	5.91	8.48
<i>İnce Memed</i> 1-4, 1955-1987	515,155	55,394	5.85	8.82

* A translation from Leonardo da Vinci by M. Belge is excluded.

For 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). Accordingly, for the volumes 1 through 4, we respectively obtained 17, 21, 31, and 32 blocks.

We obtained our style markers in the following way. For sentence length we counted the number of words in each sentence. In Turkish the number of syllables in a word is the same as the number of vowels in that word. For word length information we considered the number of characters of all words and unique words of a block.

For the selection of the “most frequent words” we considered the most frequent “context-free” 50 words of each volume and all volumes combined, and took the intersection of these 5 sets then used the members of the resultant set in our experiments (for details see Patton, Can (2004)). The list contains 33 words. These words in alphabetical order are the following (most common English meaning is provided after the word): adam (man), ben (I), beni (me), bile (even), bir (one), böyle (so), bu (this), bütün (all), çok (very), da (too), daha (more), de (too), dedi (said), diye (that), geldi (came), gibi (like), gün (day), her (every), hiç (never), kadar (until), ki (that), mi (adverb of interrogation), ne (what), onu (her, him, it), onun (hers, his, its), öyle (so), sen (you), sonra (later), şu (that), üstüne (over), uzun (long), var (there is, there are), and finally ya (then).

We obtained the stem of each word by using a statistical stemming algorithm (Altintas and Can, 2002) supported by a morphological analyzer (Oflazer, 1994). This provided the POS (part of speech) information --POS information was correct for about 80% of the cases; see Altintas, Can (2002)--. We then used the frequency of each POS as a style marker. The word types (POS

information) used are the following: adjectives, adverbs, conjunctives, determiners, duplications, interjections, nouns, numbers, post positions, pronouns, questions, and verbs (Oflazer, 1994). In the rest of the paper, we use the 3-prefix of each (e.g., “adj” for adjective) as its abbreviation.

3. Experimental Results and Discussion

We compared the style markers: token length, type length, syllable counts per word, and sentence length for changes across the four volumes. We then conducted stepwise discriminant analysis to determine the best discriminators of each style marker. Later these markers were used in a cross validation based discriminant analysis. Further investigation using multiple analysis of variance (MANOVA) revealed how the attributes of each style marker group distinguish among the volumes. All of our analyses were conducted using SAS for Windows Version 8.2.

3.1. 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 categorical). 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 II summarizes the means and standard deviations of the four markers for each of the four volumes.

Table II. Means and standard deviations for selected style markers

Volume	Token Length		Type Length		Syllable Counts		Sentence Length	
	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 I, 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) was 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. 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 II illustrates the significant increase in average sentence length from Volume 1 to Volume 4.

3.2. 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: “bile,” “bir,” “böyle,” “bu,” “çok,” “da,” “dedi,” “geldi,” “ki,” “onu,” “sen,” “şu,” “üstüne,” “var,” and “ya.” Among the syllable counts, the one, two, four, and eight syllable words provided the best separation among volumes. The best discriminators among the word types were adj, con, det, nou, num, pos, pro, que, and ver.

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 III 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% 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 III. 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%.

Table III 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 words. Using syllable counts, 60.06% of the blocks were correctly classified, etc.

3.3. MANOVA Results

To determine the volumes that were discriminated by each style marker discriminator, a 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 SL_i (SL₁, SL₂, etc.) as the number of sentences of length *i*, Table IV.a contains the average number of sentences of length *i* per block for each volume.

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

The best discriminators among the syllable counts are provided in Table IV.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 IV.d presents the average number per block of the best discriminators among the word types. In fact all twelve parts of speech served as excellent discriminators.

Table IV.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.

TableIV.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 IV.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

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

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

Table IV.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 IV.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.

In Table IV.f, TYP_i 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.

Table IV.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 V. 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	Adj, Con, Dup, Num, Ver	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	Con, Nou, Num, Ver	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	-	Adj, Nou Que	TOK3,TOK10,	-

* Empty cells are indicated by the - symbol.

** Bold sentence lengths had significant differences among all the volumes.

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 V 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.

The average number of nouns per block is significantly higher in volume 4 than volumes 1 and 3 (see Tables IV.d and V). This is in tandem with the increase in the average number of

sentences per block of length 22 and 30 between these same volumes. Possibly this is due to longer sentences having larger concentration of nouns than shorter sentences.

4. Conclusions

Our stylometric 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; 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. The literary remarks of Oğuzertem (1987) state that the first novel could be termed “romantic,” the second “realistic,” and the last two “postmodernist.” Even though our results are obtained objectively, they are consistent with this statement. The separation among volumes can also be attributed to the change of writing style with time (Can, Patton, 2003). The results of our study provide valuable information for other researchers in their stylometric investigation of agglutinative languages such as Turkish.

Acknowledgements

We thank Engin Demir, Kemal Oflazer, Süha Oğuzertem, and the Bilkent University Computer Engineering department for their support of the project at different stages.

References

- 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.
- Binongo J. N. G., Smith M. W. A. (1999) “The Application of Principal Component Analysis to Stylometry.” *Literary and Linguistic Computing*, 14(4), 445-465.
- Can, F., 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.
- Forsyth R. S., Holmes D. I. (1996) “Feature-finding for Text Classification.” *Literary and Linguistic Computing*, 11(4), 163-174.
- Hebért, E. L., Tharaud, B. (1999) *Yaşar Kemal on his Life and Art*. Syracuse University Press, Syracuse, NY.
- Oflazer, K. (1994) “Two-level Description of Turkish Morphology.” *Literary and Linguistic Computing*, 9(4), 137-149.
- Oğuzertem, S. (1987) “Yashar Kemal’s *İnce Memed*’s: Myth in the Making.” The Second Turkish Studies Conference, Indiana University, Bloomington, IN, USA.
- Patton, J. M., Can, F., (2004) “A Detailed Stylometric Investigation of the *İnce Memed* Tetralogy” Technical Report, Computer Science and Systems Analysis Department, Miami University, Oxford, OH, USA.