Determining translation invariant characteristics of James Joyce's *Dubliners*

Jon M. Patton & Fazli Can
Miami University / Bilkent University

We provide a comparative stylometric analysis of the *Dubliners* stories of James Joyce by using its original and Murat Belge's Turkish translation. We divide the stories into four categories as suggested by Belge and investigate the success of automatic classification by using discriminant analysis with various style markers. We show that different style markers show different categorization success rates and most of the style markers provide better classification rates in English. We also investigate the sentence, token and type length in both languages. We show that sentence lengths are linearly mapped from English to Turkish, type and token length distribution follow the Poisson distribution in both languages, and the related relative frequency curves provide us with an invariant between the original text and the translation.

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). (Machine learning methods are also used for data mining (Witten et al. 2011); however, in this work we only consider some statistical methods). Such studies aim to discover patterns that are 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 (Whissell 1994). Similar methods are also used in different fields that involve other kinds of human artifacts, such as architecture, music, painting, software, etc. (Sedelow 1970; Oman & Cook 1989).

In this study we analyze Joyce's *Dubliners* and its Turkish translation *Dublinliler* (Belge 2000) in order to identify translation invariant characteristics between the

source and target texts. Joyce (1882–1941) is one of the most important writers of English literature: his novel *Ulysses* has been voted the finest English-language novel published in the 20th century by a jury of scholars and writers (Lewis 1998). The translator, Murat Belge (1943 -), is a academician and prominent literary figure in Turkey. *Dubliners* was first published in 1914 and contains all the short stories of Joyce. Its (only) Turkish translation by Belge was first published in 1987. The book is regarded as the first important published work of Joyce. It contains 15 stories which are thematically connected to each other; they can be read individually or can be regarded as a part of a novel (Belge 2000: p. 5). In the preface to his translations Belge says that the mutual theme of the stories is "death in life," "being dead as being alive." He also states that the 1910s atmosphere of Ireland reflected in the stories is not too far away from that of Turkey in the 1950s. We observe that in his translation Belge tries to be loyal even to the punctuation used in the stories. Of course, here one may also recall the phrase "Traduttore, traditore" ("the translator is a betrayer") *(*Jakobson 1959*).*

As we indicated above in this work our aim is to identify style-related features of the original work which are retained in translation. In our analysis we use the style markers: (1) sentence length in terms of the number of words, (2) the most frequent words, (3) token length (word length in text), (4) type length (word length in vocabulary), and (5) type-to-token ratio (vocabulary richness). For example, the sentence "Rose is a rose is a rose is a rose." by Gertrude Stein (1922) has a length of 10 words or 10 tokens and the corresponding vocabulary contains 3 types: "a," "is," and "rose". The lengths of these types are respectively 1, 2, and 3 and the respective word frequencies are 3, 3, and 4. The type-to-token ratio of the sentence is 3/10. We investigate the sentence, token and type length in both languages, show that sentence lengths are linearly mapped from English to Turkish, the type and token length distributions follow the Poisson distribution in both languages, and the related relative frequency curves provide us with an invariant between the original text and the translation.

The rest of this chapter is organized as follows. After the next section which is on related work we provide a short discussion of Turkish language morphology. Then we describe our experimental environment and design. The section following that is devoted to experimental results obtained as we consider classifying stories using discriminant analysis, study classifying text as the English original or Turkish translation by discriminant analysis, compare sentence lengths and type-to-token ratios between the English and its Turkish translation, and study type and token relative frequency plots with the Poisson distribution. We conclude the paper by summarizing our findings and the major invariant found between the original text and its translation, and providing future research pointers.

2. Related 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 or machine learning methods. These patterns are used to resolve stylometric problems, such as authorship attribution, style change, and stylochronometry (i.e. assigning date to work). A detailed overview of the stylometry studies in literature within a historical perspective is provided by Holmes (1994). He gives a critical review of numerous style markers and reviews works on the statistical analysis of change of style with time. A solid critique of authorship studies is provided by Rudman (1997). Juola (2006), and Stamatatos (2008) provide a review of types of analysis, features, and recent developments in authorship attribution studies.

An extensively used style marker is the frequency count of "most frequent 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) in ten stylometry problems (such as authorship, chronology, subject matter, etc.) with various levels of success. The work by 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. Their study shows 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. Recent work by Popescu et al. (2011) extensively investigates the vocabulary richness problem using 1185 texts in 35 languages. In their study they investigate Turkish by using the corpus studied in (Can & Patton 2010).

The text categorization methods illustrated by Sebastiani (2002) try to assign texts into predefined categories such as known authors. Their aim is the automated categorization of texts into predefined categories as we do in this work by using discriminant analysis-based stepwise cross validation. In our previous work we studied the writing style change of two Turkish authors Yaşar Kemal and 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 the 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 that a greater time gap may have a positive impact on separation and categorization. We also have a similar study for the *İnce Memed* tetralogy of Yaşar Kemal (Patton & Can 2004). In our recent work (Can et al. 2011) we provide the first style-centered text categorization study on the Ottoman language using the poems of ten poets from five different centuries. Within the context of this language, we evaluate the performance of two different machine learning methods.

Rybicki (2006) studies two English translations of a well-known trilogy of the Polish author Henryk Sienkiewicz. For this purpose he uses a multidimensional scaling technique and shows that novel character idiolects are preserved in translations. In our recent work (Can et al. 2010) we use clustering to show that patterns in source texts reappear in translations. For this purpose we use Shakespeare's sonnets and their Turkish translations (in unreported additional experiments we show the same in German and Latin translations). A translation must retain the original meaning of the source text. In (Can et al. 2010) we captured this in terms of the presence of parallel clustering structures in source and target texts. If retained these parallel structures can be also be regarded as a translation invariant characteristic. In another recent work we (Altintas et al. 2007) introduce a method called PARTEX-M that uses time-separated parallel translations to quantify diachronic changes in a target language. We show that over time, for both text and lexicon, the length of Turkish words has become significantly longer, word stems have become significantly shorter, and the vocabulary richness of the language has dropped when measured as type-to-token ratio using word stems. Pado et al. (2009) propose a machine translation evaluation metric that considers the semantic equivalence of the translation to its original. They use the Stanford Entailment Recognizer and set a textual entailment between the original and translated texts which contains "common sense" reasoning patterns that is used to hold a relationship between the languages.

3. Turkish language morphology

Turkish belongs to the Altaic branch of the Ural-Altaic family of languages. Its alphabet, in its current orthography, adopted in 1928, is based on Latin characters and has 29 letters, eight vowels: { a, e, ı, i, o, ö, u, ü} and 21 consonants: { b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z}. In some words borrowed from 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.

Turkish is a free constituent order language, i.e. according to text flow and discourse context at certain phrase levels; its constituents can change order

(Lewis 1988). It is an agglutinative language similar to Finnish and Hungarian. Such languages carry syntactic relations between words or concepts through discrete suffixes and have complex word structures. Turkish words are constructed using inflectional and derivation suffixes, and inflectional and derivational morphemes are linked to a root.

Consider the following example.

Okutamayacakmışçasına (*oku* + *t* + *ama* + *yacak* + *mış* + *çasına*)
(as if not being able to make [*them*] read)

The root of this example is verb *oku* (to read) and the meaning is enriched through affixation of derivational and inflectional suffixes. Verbs can be converted into nouns and other forms as well as nouns can be converted into verbs and other grammatical constructs through affixation. For further information and examples see (Lewis 1988).

In contemporary everyday Turkish, it is observed that words have about 3 to 4 morphemes including the stem with an average of 1.7 derivations per word (Oflazer 2003). 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 word stem, it is possible to 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: 31). Like other agglutinative languages, in Turkish it is possible to have words that would be translated into a complete sentence in non-agglutinative languages such as English.

Studies of Turkish morphology as a computation problem include (Köksal 1973; Solak & Oflazer 1993). A two-level (lexical & 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).

4. Experimental environment and design

4.1 Experimental environment

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* is the set of all types.

Table 1 shows the *Dubliners* stories divided into four categories. We use the classification suggested by Belge (2000): the first three stories are related to children, the next four are related to young people, the next four are related to

Table 1. *Dubliners* stories and style marker overall characteristics

Cat.	Story[1]	English					Turkish					
		N	V	Avg. TL	Avg. YL	Avg. Sen. Len.	N	V	Avg. TL	Avg. YL	Avg. Sen. Len.	Nt/ Ne
1	*The Sisters*	3109	899	4.074	5.828	15.02	2071	1256	6.092	7.261	9.678	.666
	An Encounter	3251	981	4.127	5.909	17.29	2129	1313	6.135	7.166	11.16	.655
	Araby	2343	823	4.108	5.763	16.05	1486	1048	6.509	7.365	10.11	.634
2	*Eveline*	1831	628	4.128	5.573	14.19	1176	805	6.242	7.117	8.842	.642
	After the Race	2241	859	4.469	6.146	15.67	1613	1064	6.434	7.432	11.28	.720
	Two Gallants	3923	1145	4.201	5.86	12.78	2626	1529	6.257	7.314	8.444	.669
	The Boarding House	2813	941	4.197	5.832	16.74	1862	1226	6.308	7.29	10.89	.662
3	*A Little Cloud*	4936	1379	4.276	5.906	11.95	3405	1904	6.088	7.24	8.305	.690
	Counterparts	4108	1102	4.213	5.897	14.21	2704	1556	6.23	7.262	9.135	.658
	Clay	2657	719	4.038	5.437	23.1	1673	995	6.079	7.116	12.39	.630
	A Painful Case	3640	1233	4.437	6.247	17.33	2572	1617	6.404	7.454	12.19	.707

(*Continued*)

Table 1. (continued)

4	*Ivy Day in the Committee Room*	5231	1241	4.016	5.789	10.14	3700	1789	5.61	7.003	7.034	.707
	A Mother	4541	1189	4.388	6.139	15.77	3228	1761	6.271	7.547	10.94	.711
	Grace	7526	1791	4.282	6.381	11.63	5533	2672	5.898	7.436	8.137	.735
	The Dead	15709	2744	4.229	6.33	15.55	11211	4676	6.04	7.712	10.85	.714
–	All stories	67859	7345	4.221	6.85	14.21	46989	14470	6.104	8.221	9.598	.693

* N: no. of tokens, V: no. of types, Avg TL: Average Token Length, Avg YL: Avg. type length, Avg. Sen. Len.: Average Sentence LengthN_t/N_e: Number of Turkish Tokens/ Number of English Tokens.

1. The names of the stories in Turkish in the same order are as follows: *Kızkardeşler, Bir Karsılaşma, Araby, Eveline, Yarıştan Sonra, İki Çapkın, Pansiyon, Küçük Bir Bulut, Suretler, Toprak, Üzücü Bir Olay, İdarehanede Ulusal Bayram Günü, Bir Anne, Arınma, Ölüler.*

middle-aged people, and finally the last four stories are on social events. The table lists for each story the total number of tokens, the total number of types, the average token length, the average type length, and the average sentence length for both the English original and the Turkish translation. The last column provides the ratio between the number of tokens in the Turkish translation and the English original. A sample from the original and the translation is provided in Figure 1.

The Sisters
There was no hope for him this time: it was the third stroke. Night after night I had passed the house (it was vacation time) and studied the lighted square of window: and night after night I had found it lighted in the same way, faintly and evenly. If he was dead, I thought, I would see the reflection of candles on the darkened blind for I knew that two candles must be set at the head of a corpse. He had often said to me: "I am not long for this world," and I had thought his words idle. Now I knew they were true.

Kızkardeşler
Bu sefer hiç umut kalmamıştı: üçüncü krizdi. Üstüste birkaç gece evinin önünden geçmiş (tatildeydik o sıra), pencerenin aydınlık dikdörtgenini gözlemiştim: her seferinde aynı şekilde aydınlandığını görüyordum, hafif ve dengeli bir ışıkla. Ölmüş olsa, diye düşünüyordum, kararık perdenin üzerinde mumların yansımasını görmem gerekirdi, çünkü ölülerin başucuna iki mum dikildiğini biliyordum. Kendisi de sık sık sözünü ediyordu, "Bu dünyada çok günüm kalmadı artık" diye; oysa ben bunları laf olsun diye söylenmiş sözler sanmıştım. Doğru olduklarını şimdi anlıyordum.

Figure 1. An excerpt of text taken from the beginning of the short story "*The Sisters*" followed by the Turkish translation (Belge 2000) "*Kızkardeşler*"

Note the overall ratio for the entire stories provided at the bottom of the last column is .6925. Since we use blocks of text in many of our analyses, we use the factor .6925 in determining the block size of Turkish translation text based on the English block size. The English block size was set to 2500 words. Thus the Turkish block size was set to .6925 * 2500 or 1732 (rounded to next higher integer) words. This results in 27 blocks for each language.

4.2 Experimental design

As indicated earlier we used five style markers (1) sentence length in terms of number of words, (2) most frequent words, (3) word length in text, (4) word length in vocabulary, and (5) type-token ratio (vocabulary richness).

We use the 60 most frequent words of the English original and the Turkish translation (see Figure 2). Our experience shows that 60 words provide us with a better selection of discriminators than a smaller number. 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. For our first set of discriminant analysis experiments, we used sentences having lengths up to 47 words from the original English stories and up to 37 words from the Turkish translations as potential discriminators.

a, all, an, and, as, at, be, been, but, by, down, for, from, had, have, he, her, him, his, I, if, in, into, is, it, little, man, me, my, no, not, of, old, on, one, or, out, said, she, so, some, that, the, their, them, then, there, they, to, up, was, we, were, what, when, which, who, with, would, you

adam (man), *ama* (but), *bana* (to me), *başladı* (started), *ben* (I), *bir* (one), *biraz* (a little), *birkaç* (a few), *böyle* (so), *bu* (this), *bütün* (all), *büyük* (large), *çok* (very), *çünkü* (because), *da* (too), *daha* (more), *de* (too), *dedi* (said), *değil* (not), *diye* (that), *doğru* (right), *en* (most), *etti* (did), *evet* (yes), *genç* (young), *gibi* (like), *güzel* (fine), *her* (every), *hiç* (never), *için* (for), *içinde* (in), *iki* (two), *ile* (and), *iyi* (good), *kadar* (until), *kendi* (herself, himself), *ki* (that), *küçük* (small), *mi* (adverb of interrogation), *nasıl* (how), *ne* (what), *o* (he, she, it), *olan* (being), *olarak* (happening), *olduğunu* (is), *ona* (to her, to him), *onu* (her, him, it), *onun* (hers, his, its), *öyle* (so), *sonra* (later), *sordu* (asked), *söyledi* (said), *şey* (thing), *şimdi* (now), *vardı* (there was, there were), *ve* (and), *ya* (then, so), *yeniden* (again), *yok* (there is no), *zaman* (time)

Figure 2. The most frequent 60 English and Turkish words listed in alphabetical order. Each Turkish word is followed by its English translation (words may have more similar meanings). The words "*da*" and "*de*" are essentially the same word; the surface difference is due to the vowel harmony requirement in Turkish

For word length information we considered the number of characters of all words and unique words of a block. We used English words having up to 16 characters and Turkish words up to 20 characters for our first set of discriminant analysis experiments.

4.3 Use of vocabulary richness in discriminant analysis

The type-token ratio measures the richness of the vocabulary for a block of text. Figure 3 depicts the plot of type-token ratio for both the English original and the Turkish translation of the *Dubliners* versus percent of text.

The plot illustrates that the ratio is considerably higher for the Turkish translation then the English original text. Also for both curves, the ratio decreases as the percent of text increases. Thus in order to make comparisons involving the English and Turkish versions, the same block size needs to be used for each. We decided on a block size of 1150 words which is the approximate number of tokens in the Turkish translation of *Eveline* (see Table 1).

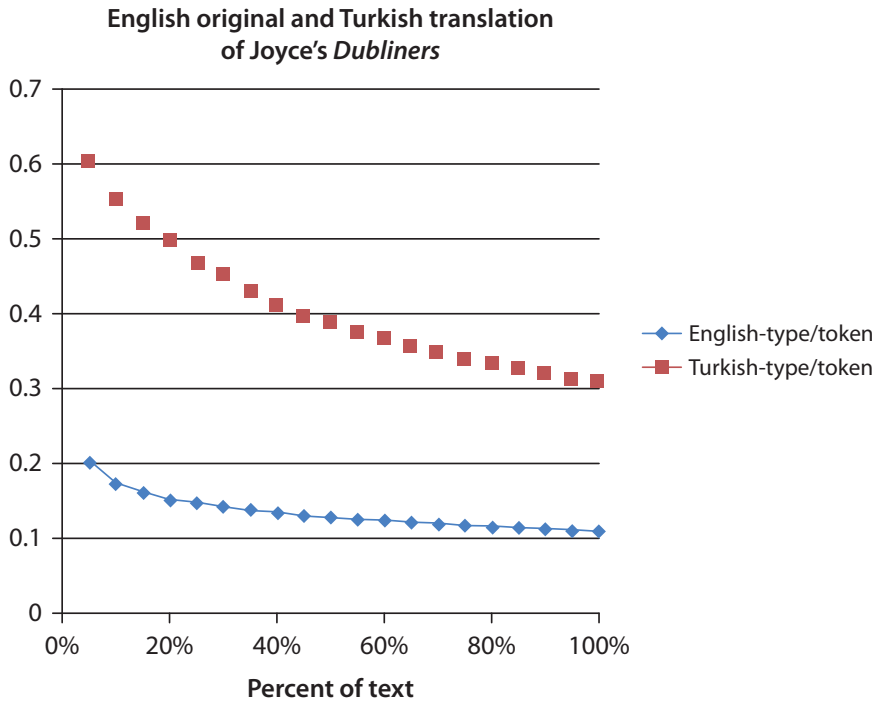

Figure 3. Type-to-token ratio plot of English and Turkish version of *Dubliners*

5. Experimental results and discussion

5.1 Classifying stories: Discriminant analysis results

Discriminant analysis is a statistical technique that uses the information available in a set of independent variables to predict the value of a categorical dependent variable. Usually this dependent variable is coded as a series of integer values that represent the categories to which the observations belong. The goal of discriminant analysis is to develop a classification rule for predicting the category to which a new observation will most likely belong based on the values the independent variables. The rule is based on values of discriminant function(s) developed by the procedure. These function(s) are linear combinations of values of the independent variables and are optimized to provide the best classification rate for the dependent variables in the sample.

A stepwise discriminant analysis is often conducted beforehand to select a good set of discriminators among the pool of independent variables. During each step, one variable, currently not in the model, is chosen as a candidate to enter the model if it provides the largest contribution to the discriminatory power of the model. If this contribution meets a criterion set by the user, it enters the model. Then the variable in the model with the the smallest discriminatory power is

examined. If its discriminatory power does not meet the criterion to stay in the model, it is removed from the model. The process ends when variables not in the model do not meet the criterion to enter, and all the variables in the model meet the criterion to stay.

We used discriminant analysis to classify the stories into their respective categories as defined previously. Separate discriminant analyses were done in both English and Turkish using the frequencies of each of the following style markers: most frequent words, token length, type length and sentence length. To classify a story into a category using frequencies of a style marker, a stepwise discriminant analysis was conducted on the other stories to determine the best discriminators for that category. Using these discriminators, an additional discriminant analysis was conducted to classify this story using cross validation. In cross validation each story in turn is excluded from the rest of the stories in the derivation of linear discriminant functions employed for classifying each story in one of the four categories. Then the excluded story is classified by these linear discriminant functions. This two stage process employs a strict form of cross-validation that completely eliminates bias from the story to be classified. When type-token ratio is the only discriminator, we bypass the stepwise discriminant analysis stage and only conduct the cross validation procedure. All of our analyses were conducted using SAS for Windows, Version 9.2.

Since each story has a different number of words and sentences, we used relative frequencies for each style marker instead of the actual frequencies. For example, the number of sentences in a story having a certain length was divided by the total number of sentences in that story in order to get the relative frequency of that sentence length. A similar process was applied to tokens, types, and the most frequent words. This serves to eliminate any bias due to the length of the story. When type-token ratio was the only discriminator, we used the first block of 1150 words from each story.

Overall, the best discriminators among the best English words in the *Dubliners* are "I", "had", "as," "down," "you," "out," "up," "little," "his," "so," "be," "all," and "one." The best discriminators among the most frequent Turkish words in the translations were the following: "*sordu*," "*onun*," "*ben*," "*ya*," "*kadar*," "*diye*," "*onu*," "*hiç*," "*etti*," "*biraz*," "*çok*," and "*sonra*." For token lengths, the words of length 1 were the best discriminators for the English original; whereas, words of length 2, 6, 13, and 20 were the best for the Turkish translation. For type lengths, vocabulary words of length 1, 2, and 5 were the best for the English, and vocabulary words of length 3, 12, and 19 were best for the Turkish translation. For sentence lengths, sentences having 9, 23, 37, and 38 words provided the best separation for the English version, whereas sentences of length 2, 4, 7, 14, and 16 words were best for the Turkish.

Table 2 summarizes the series of discriminant analyses performed on the stories. Each block in the table indicates the percentages of stories taken from the story group given by the row and column header. The number in parentheses following the row header is the number of stories in that group. The first row in each block contains the percentage of correct classification using discriminators

Table 2. Correct classification rates of story categories for each style marker

Category	Style marker	Children		Young adults		Middle age		Social	
		Engl.	Turk.	Engl.	Turk.	Engl.	Turk.	Engl.	Turk.
Children (3)	Sentence Length	0%	0%	33%	67%	33%	0%	33%	33%
	Most Frequent Words	67%	0%	0%	33%	0%	33%	33%	33%
	Token Lengths	33%	33%	0%	33%	0%	33%	67%	0%
	Type Lengths	33%	0%	0%	33%	33%	67%	33%	0%
	Type-Token Ratio	0%	67%	0%	33%	67%	0%	33%	0%
Young Adults (4)	Sentence Length	25%	0%	25%	0%	0%	25%	50%	75%
	Most Frequent Words	0%	25%	50%	25%	0%	25%	50%	25%
	Token Lengths	0%	0%	50%	25%	25%	50%	25%	25%
	Type Lengths	25%	0%	75%	0%	0%	100%	0%	0%
	Type-Token Ratio	50%	0%	0%	0%	25%	75%	25%	25%
Middle Age (4)	Sentence Length	0%	0%	25%	25%	25%	50%	50%	25%
	Most Frequent Words	0%	0%	0%	75%	50%	0%	50%	25%
	Token Lengths	0%	25%	50%	50%	50%	25%	0%	0%
	Type Lengths	75%	0%	0%	50%	25%	50%	0%	0%
	Type-Token Ratio	50%	50%	0%	0%	0%	0%	50%	50%
Social (4)	Sentence Length	0%	75%	25%	25%	50%	0%	25%	0%
	Most Frequent Words	0%	0%	0%	50%	0%	0%	100%	50%
	Token Lengths	25%	0%	25%	0%	25%	75%	25%	25%
	Type Lengths	25%	25%	0%	25%	0%	0%	75%	50%
	Type-Token Ratio	50%	0%	0%	25%	0%	25%	50%	50%

Overall classification rate for sentence length: 20% (3/15: 3 out of 15) for English and 13.33% (2/15) for Turkish, most frequent words: 66.67% (10/15) for English and 20% (3/15) for Turkish, token lengths 40% (6/15) for English and 26.67% (4/15) for Turkish, type lengths: 53.33% (8/15) for English and 26.67% (4/15) for Turkish, type-token ratio: 13.3% (2/15) for English and 26.7% (4/15) for Turkish.

based on sentence length. The second, third, fourth, and fifth row in each block contains the percentage of correct classification based respectively on the frequencies of the most frequent words, token lengths, type lengths, and type-token ratio.

Two columns are associated with each column header; one contains the classification rate for the English version, the other for the Turkish translation. For example, the (numerical data) block in the last row and last column of the table indicates that, of the 4 stories in the Social story group, 25% (1 out of the 4) of the English originals were correctly classified as belonging to this group based on sentence length. In the next column 0% (0 out of the 4) of the Turkish translations were correctly classified. Of the 75% of the English stories that were incorrectly classified, 50% or two stories were classified in the Middle Age story group and one was classified in the Young Adults group.

For the frequencies of the most frequent words 100% of the English stories in the Social group were correctly classified and 50% of the translations were correctly classified. Likewise, using the token length discriminators, 25% of the English stories and 25% of the Turkish translations were correctly classified. Using the type length discriminators, 75% of the English stories and 50% of the Turkish translations were correctly classified. Finally, using type-token ratio, 50% of English stories and 50% of the Turkish translations were correctly classified.

The footnote at the bottom of Table 2 displays the overall correct classification rates for each of the attributes. 67 % of the English stories and 20% of the corresponding Turkish translations were correctly classified using the frequency of the most frequent context free words. Using sentence lengths, 20% of the English stories were correctly classified as well as 13.3% of the Turkish translations. Using type lengths, we had 53.3% correct classification for English and 26.7% for Turkish. Using type-token ratios, we had 13.3% for English and 26.7% for Turkish.

Overall, the most frequent words was the style marker having the best classification rate for the English stories at 66.67%. This is excellent considering the strict cross validation procedure used to classify the stories. For the Turkish translations the best style markers for classification were token lengths, type lengths, and type-token ratios each having rates of 26.7%, which is only slightly better than chance. In general for each style marker, except for type-token ratio, the Turkish translations had a lower classification rate than the English stories. Thus it appears that certain nuances in the English originals, necessary for discrimination, are lost in the Turkish translation.

5.2 Classifying text as the English original or Turkish translation

Using the first 1150 words in each of the 15 stories for both English and Turkish versions, we did a discriminant analysis using type-token ratio as the only

discriminator in classifying whether the story was either the English original or the Turkish translation. As one might surmise, the classification rate using cross-validation was 100%.

Table 3 shows this result as well as the classification rates of the other style markers. For these other style markers we used blocks of 2500 words from the complete original and of 1732 words from the translated stories. As a result we used 27 blocks from both the originals and the translations for classification. As in the previous discriminant analysis, we used the strict form of cross validation in order to eliminate any bias.

Table 3. Classification of blocks according to their language

Group	Style marker	English	Turkish
English	Sentence Length	67%	33%
	Most Frequent Words	100%	0%
	Token Lengths	100%	0%
	Type Lengths	100%	0%
	Type-Token Ratio	100%	0%
Turkish	Sentence Length	0%	100%
	Most Frequent Words	0%	100%
	Token Lengths	0%	100%
	Type Lengths	0%	100%
	Type-Token Ratio	0%	100%

5.3 Comparison of sentence lengths and type-token ratios between English and Turkish translation

The classification results indicated that all style markers except sentence length had perfect classification rates. We would have expected this for type-token ratio and Most Frequent Words, but it is not as obvious for token and type lengths. For sentence length all of the Turkish blocks and 67% of the English blocks were classified correctly. Considering the strict cross validation procedure used in this discriminant analysis, 67% is quite good!

5.4 Linear relationship of sentence length between English and Turkish translation

We next investigate the possible linear relationship of sentence length between the English and Turkish translation, Figure 4 provides a scatterplot of average Turkish sentence length vs. English sentence length for each block of the *Dubliners*.

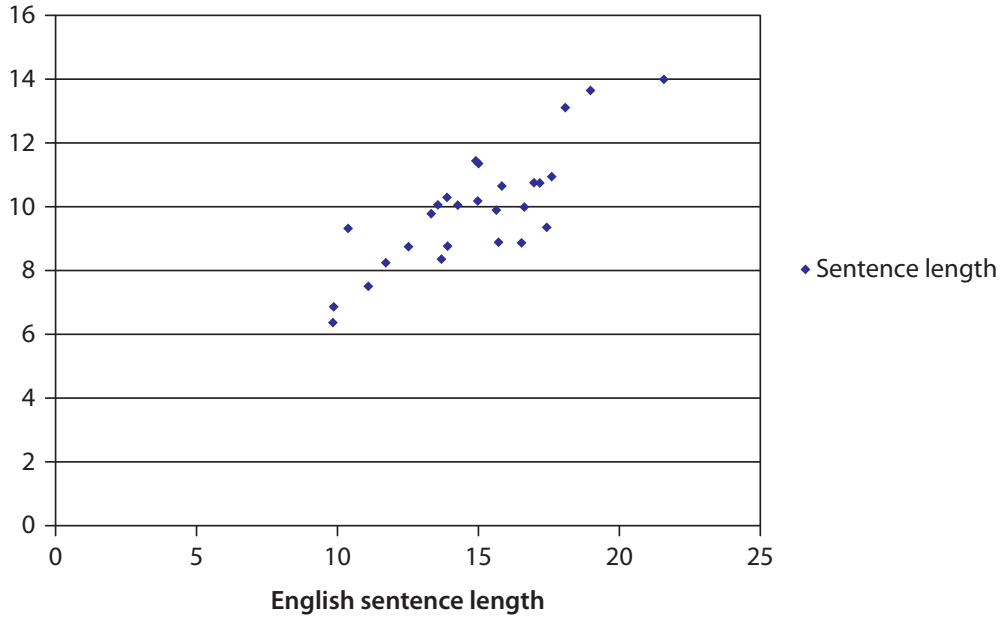

Figure 4. Plot of average Turkish sentence length vs English sentence length for each block

A regression analysis was performed using average Turkish sentence length as the response variable and average English sentence length as the predictor. A no intercept model was attempted since an English sentence of zero length corresponds to a translated Turkish sentence of zero length. The regression results gave an extremely strong relationship between English and Turkish sentence lengths ($F(1, 26) = 46.2$ prob value $<.001$). The coefficient of determination (R^2) was .640 indicating that 64% of the total variance of English sentence length about its average can be explained by the model. The prediction equation is given by

Average Turkish sentence length = .666 * English sentence length.

This equation can be interpreted as follows: for every extra word that is added to an English sentence, on the average the translated Turkish sentence increases by approximately 2/3 of a word.

To compare the type-token ratios between the English and Turkish translations, 1150 word blocks were used. The average ratio for the English blocks was .469, and for the Turkish translation, it was .703. The ratio of the English average over the Turkish average is .469/.703 = .667.

5.5 Comparison of type and token relative frequency plots with the Poisson distribution

A Poisson distribution describes a discrete random variable representing the number of occurrences of an event over a specified interval of time or space.

For example such an event might be the number of customers who arrive at the checkout counters in one hour.

Another event would be the number of times a type is used in a block of 1000 tokens.

It has been known for some time that the distribution of relative frequency plots of token and type lengths can be approximated by a Poisson distribution (Baayen 2001). Figures 5 and 6 display relative frequency plots; one graph displays token lengths of the English original and Turkish translation of the *Dubliners*, whereas the other graph displays type lengths. Superimposed on each of these plots are two Poisson distribution curves. On the token length relative frequency graph, the mean of one of the Poisson distributions is the mean token length of the English original. The mean of the other is the mean token length of the Turkish translation. On the type length graph, each Poisson distribution curve has a mean that is equal to the type mean of the corresponding English and Turkish translation.

The Komolgorov-Smirnov statistic can be used to test whether two relative frequency plots have similar shapes or are based on a sample from a given distribution (such as the normal or Poisson distribution). The statistic is largely based on the squared vertical distance between the points on the curves for each type or token length.

A series of two sample Komolgorov-Smirnov tests were conducted to determine whether the relative frequency plot of the token lengths of the original English version of *Dubliners* has a different distribution than the corresponding plot of the Turkish translation. We were also interested in whether the distributions of the English and Turkish token relative frequencies are different to their corresponding Poisson distributions.

Table 4. Results of the Kolmolgorov-Smirnov tests

Token length				Type length			
Group 1	**Group 2**	**Ksa**	**Pr > Ksa**	**Group 1**	**Group 2**	**Ksa**	**Pr > Ksa**
English	Poisson English	.343	.9996	English	Poisson English	.514	.954
Turkish	Poisson Turk.	.474	.9780	Turkish	Poisson Turk.	.316	1.000
English	Turkish	.791	.560	English	Turkish	.632	.819

We tested a similar set of hypothesis using type lengths. Table 4 summarizes the results of these tests. Poisson_English refers to the Poisson distribution whose mean is the same as the English token or type length mean (depending on the table). Likewise Poisson_Turkish refers to the Turkish translation token or type

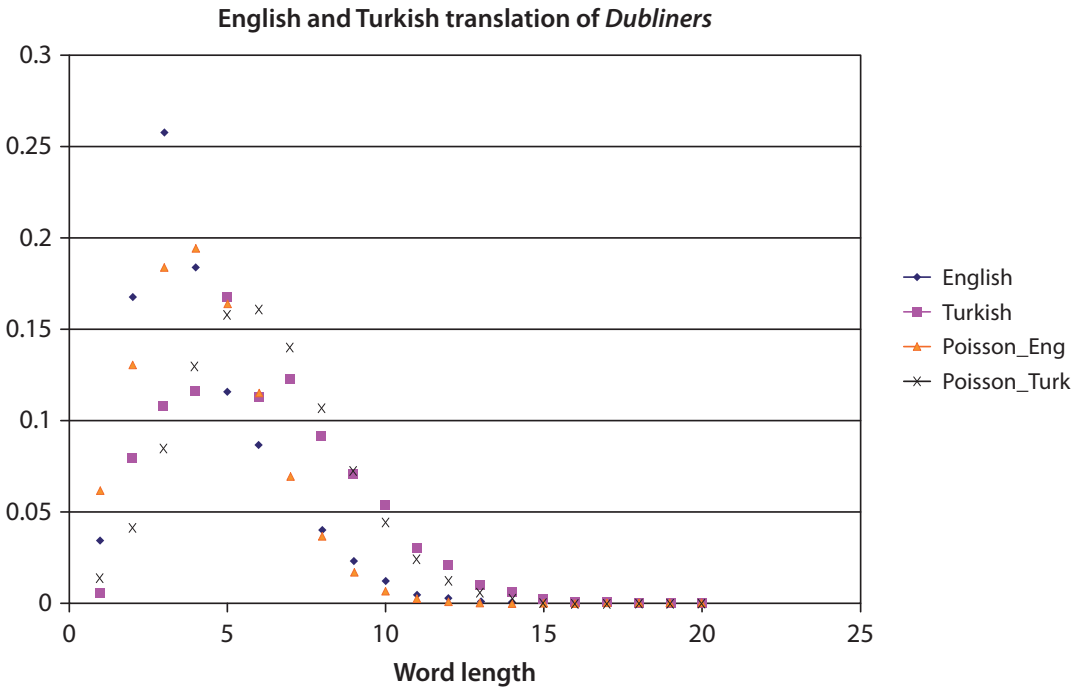

Figure 5. Relative frequency token (word) length and Poisson distribution plots of *Dubliners*

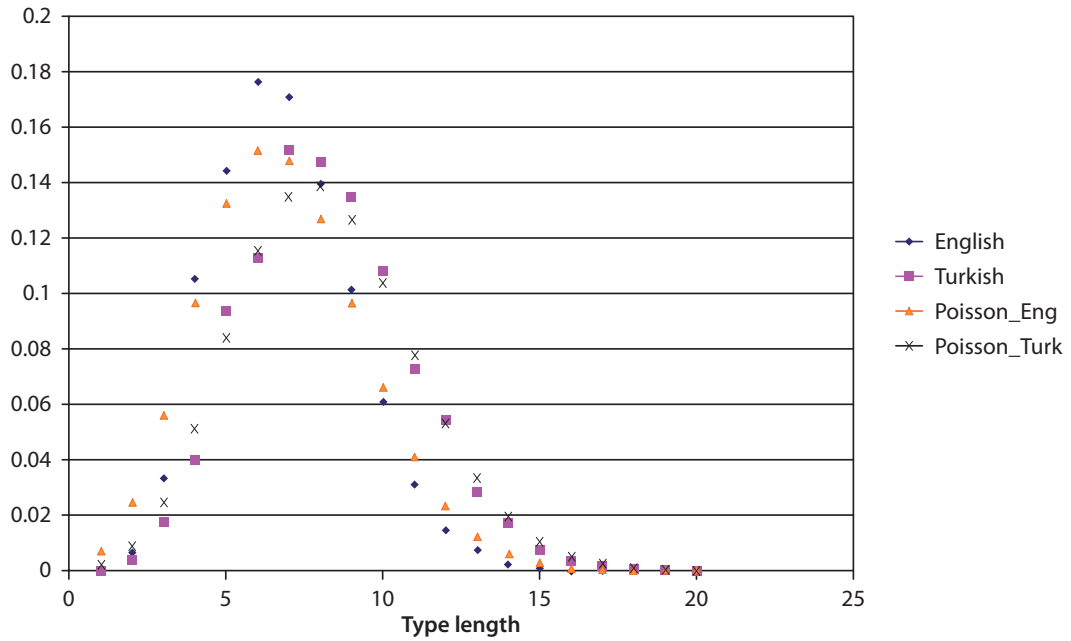

Figure 6. Relative frequency type length and Poisson distribution plots of *Dubliners*

mean. Ksa is the asymptotic Komolgorov-Smirnov statistic. The column header Pr > Ksa is the p-value corresponding to the null hypothesis that the two curves in question come from the same distribution.

All of the p-values are very high; none are less than .05. There is insufficient evidence to reject the hypotheses that

1. The relative frequency curve of the token (type) lengths of the English version of the *Dubliners* has the same distribution as the Poisson distribution having the same mean as that of the English token (type) length.
2. The relative frequency curve of the token (type) lengths of the Turkish version of the *Dubliners* has the same distribution as the Poisson distribution having the same mean as that of the Turkish token (type) length.
3. The relative frequency curve of the token (type) lengths of the English version of the *Dubliners* has the same distribution as the relative frequency curve of the token (type) lengths of the Turkish translations.

The findings from these tests provide us with an invariant between the English original text and the Turkish translation.

6. Summary and conclusions

In this study we provide a comparative stylometric analysis of James Joyce's *Dubliners* using the original work and Turkish translation with five style markers: "sentence length in terms of words," "the most frequent words," "token length," "type length," and "type-token ratio." We categorize 15 stories of *Dubliners* into their (so called) respective groups using a combination of stepwise discriminant analysis and linear discriminant analysis using cross validation. In this endeavor our aim is to see the classification success rate with these style markers.

Our investigation shows that the same style markers can show different success rates in different languages for classifying the same information expressed in different languages. Our experiments indicate that in a majority of the cases, the nature (category) of the stories is better reflected by the originals. Only the type-to-token ratio indicated a somewhat improved classification rate in the Turkish translations.

The style marker that best classified the English stories was the most frequent words. Its classification rate at 67% is very good considering the strict cross validation process used in the discriminant analysis. On the other hand the worst style markers for classification were sentence length and type-to-token ratio. These may be qualities developed through the author's writing experiences and may remain invariant relative to the type of stories being written.

Outstanding discriminant results were achieved when the classification criteria was whether the block came from the English original or the Turkish translation. We got 100% correct classification results for each of the four style

markers; token length, type length, most frequent words, and type-to-token ratio. Only the style marker sentence length provided less than 100% perfect classification when English blocks were being classified. However, 67% (18 out of 27 blocks correctly classified) is still substantially better than chance.

Inspecting the columns of Table 1 and the plots of Figures 3 and 4 provides us with additional understanding regarding these strong results. In Table 1 the average English token length is less than the average Turkish token length for each story. We get the same observation comparing the average type length columns. Figure 3 displays a plot of the type-to-token ratio for both the English originals and Turkish translation of *Dubliners* vs percent of text. We notice that the Turkish ratio is consistently higher than the English for each percentage. Using blocks of the first 1150 words in each story, we found the ratio of the English type-to-token ratio over the Turkish ratio to be .667.

Figure 4 displays a plot indicating a strong linear relationship between average sentence length in words of a Turkish block with that of the corresponding English block. When a no-intercept linear regression line was fitted between the points, we found that the ratio of the average Turkish sentence length over the average English sentence length was approximately .666.

The major invariant found between the original text and its translation was the relative frequency plot of both token and type lengths. Since both of these plots can be modeled using the Poisson distribution, we compared the relative frequency plot of token lengths with the Poisson distribution having the same word length mean. We also compared a similar series of plots for the Turkish translations. Using non-parametric goodness of fit tests, we found that there was no significant difference between the relative frequency plots of the English and Turkish translations as well as no difference between each language plot and their corresponding Poisson distribution plots. We got similar results comparing type lengths instead of token lengths.

The results of this study can be used to check the consistency between a work and its translation. The results may also be used to diagnose plagiarism, where the potential plagiarized copy can be assumed to be a translation of the original.

References

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

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. Harald, Van Halteren Hans & Tweedie Fiona J. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3): 121–131.

Belge Murat. 2000. *Dublinliler*, 8th edn. İstanbul: İletişim Yayınları.

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

Can, Ethem Fatih, Can, Fazli, Duygulu, Pinar & Kalpakli, Mehmet, 2011. Automatic categorization of Ottoman literary texts by poet and time period. ISCIS 2010: 117–120.

Can, Fazli, Can, Ethem Fatih & Karbeyaz Ceyhun. 2010 Translation relationship quantification: A cluster-based approach and its application to Shakespeare's sonnets. ISCIS 2010: 117–120.

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

Can, Fazli & Patton, Jon M. 2010. Change of word characteristics in 20th century Turkish literature: A statistical analysis. *Journal of Quantitative Linguistics* 17(3): 167–190.

Fayyad, Usama, Piatetsky-Shapiro, Gregory & Smyth Padhraic. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11): 27–34.

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

Hakkani-Tür, Dilek Z. 2000. Statistical Language Modeling for Agglutinative Languages. PhD dissertation, Bilkent University.

Holmes, David I. 1985. The analysis of literary style - a review. *Journal of the Royal Statistical Society, Series A* 148(4): 328–341.

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

Jakobson, Roman 1959. On linguistic aspects of translation. In *On Translation,* Reuben Brower (ed.), 232–239. Cambridge MA: Harvard University Press.

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

Köksal, Aydın. 1973. Automatic Morphological Analysis of Turkish. PhD dissertation, Hacettepe University.

Lewis, Geoffrey L. 1988. *Turkish Grammar*, 2nd edn. Oxford: OUP.

Lewis, Paul. 1998. '*Ulysses*' on Top Among 100 Best Novels. New York Times, July 20, 1998. ⟨http://partners.nytimes.com/library/books/072098best-novels.html⟩ (11 August 2011).

Oflazer Kemal. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing* 9(2): 137–148.

Oflazer Kemal. 2003. Dependency parsing with an extended finite-state approach. *Computational Linguistics* 29(4): 515–544.

Oman, Paul W. & Cook, Curtis R. 1989. Programming style authorship analysis. In Proceedings of the 17th Annual ACM Computer Science Conference, 320–326.

Padó, Sebastian, Cer, Daniel, Galley, Michel, Jurafsky, Dan & Manning, Christopher D. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation* 23(2–3): 181–193.

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

Popescu, Ioan-Iovitz, Čech, Radek & Altman, Gabriel. 2011. *The Lambda-structure of Text*. Lüdenscheid: RAM-Verlag.

Rudman, Joseph. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31(4): 351–365.

Rybicki, Jan. 2006. Burrowing into translation: Character idiolects in Henryk Sienkiewicz's trilogy and its two English translations. *Literary & Linguistic Computing* 21(1): 91–103.

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

Sedelow, Sally Yeates. 1970. The computer in the humanities and fine arts. *ACM Computing Surveys* 2(2): 89–110.

Solak, Aysin & Oflazer, Kemal 1993. Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing* 8(3): 113–130.

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

Stein, Gertrude. 1922. *Geography and Plays*. Madison WI: Univ of Wisconsin Press.

Whissell, Cynthia 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.

Witten, Ian H., Frank, Eibe & Hall, Mark A. 2011. *Data Mining Practical Machine Learning Tools and Techniques*, 3rd edn. Burlington MA: Morgan Kaufmann.