Language Change Quantification Using Time-separated Parallel Translations

Kemal Altintas[a]
Computer Science Department, University of California, Irvine, Irvine, CA 92612, USA

Fazli Can[a]
Computer Science and Systems Analysis Department, Miami University, Oxford, OH 45056, USA

Jon M. Patton[a]
Information Technology Services, Miami University, Oxford, OH 45056, USA

Abstract

We introduce a systematic approach to language change quantification by studying unconsciously used language features in time-separated parallel translations. For this purpose, we use objective style markers such as vocabulary richness and lengths of words, word stems and suffixes, and employ statistical methods to measure their changes over time. In this study, we focus on the change in Turkish in the second half of the twentieth century. To obtain word stems, we first introduce various stemming techniques and show that they are highly effective. Our statistical analyses show that over time, for both text and lexicon, the length of Turkish words has become significantly longer, and word stems have become significantly shorter. We also show that suffix lengths have become significantly longer for types and the vocabulary richness based on word stems has shrunk significantly. These observations indicate that in contemporary Turkish one would use more suffixes to compensate for the fewer stems to preserve the expressive power of the language at the same level. Our approach can be adapted for quantifying the change in other languages.

Correspondence:
Fazli Can, Computer Engineering Department, Bilkent University, Bilkent, Ankara 06800, Turkey.
E-mail:
canf@cs.bilkent.edu.tr

1 Introduction

The change in natural languages is a never-ending process (Aitchison, 2001). Language changes include grammar, most frequent words, pronunciation, vocabulary, word order, word length, etc. Our aim in this study is to introduce an approach that quantifies the change by examining some unconsciously used language features (e.g. vocabulary richness and lengths of words, word stems, and suffixes). We demonstrate that the language change can be quantified by examining such language

[a]All authors contributed equally to this work and are listed in alphabetical order.

Literary and Linguistic Computing, Vol. 22, No. 4, 2007.
doi:10.1093/llc/fqm026 Advance Access Published on 21 September 2007

features in time-separated parallel translations using statistical methods. Since our language change measurement approach is based on parallel old and new texts, we refer to it as PARTEX-M (pronouced 'partexem'): 'PARallel TEXt-based language change measurement Method.' In this study, we focus on the Turkish language, specifically Turkish used in Turkey whose 'diachronic' change in the twentieth century is easily recognizable (Lewis, 1999), but has never been quantified.

Language change can be attributed to many different causes (Aitchison, 2001; Holt, 2003). In Turkish it can, at least partly, be attributed to the official state policies which aimed to eliminate the Arabic and Persian grammatical features from the language (Lewis, 1999). Nonetheless, Turkey is not the only nation that has had an experience like this (Lewis, 1999; Carroll, 2001).

We employ our PARTEX-M approach to study the Turkish language change in approximately the second half of the twentieth century. We use old and new Turkish translations of various literary works in three different (source) languages. The average time gap between old and new translations is slightly more than fifty years.

In this study, the term *word* indicates any sequence of characters that begins with a letter and continues with a letter, a number or an apostrophe sign, and a sequence of one or more characters. We use the term *token* to mean a word occurring in a given text and the term *type* to mean a word occurring in the list of distinct words (vocabulary).

In Turkish, it is possible to generate several words from a stem due to its agglutinative nature. It would be inaccurate to measure its change by only examining tokens and types as they appear in the text in their surface forms. Therefore, we develop effective stemming tools for Turkish and employ one of them in quantifying changes in Turkish. Our study shows little difference in terms of number of tokens used in old and new translations. However, we show that the stem level vocabulary richness; measured by type-to-token ratio, TTR, (no. of types)/(no. of tokens), has changed. A series of discriminant analysis experiments shows that the old and new translations are mostly distinguishable from each other when token and type lengths are used. By regression analysis, we show that longer tokens and types tend to come from new translations. We further quantify the language change by additional statistical experiments and show that suffixes are longer and stems are shorter in new translations.

The rest of the article is organized as follows. In Section 2, we give an overview of previous work on language change. A describtion of PARTEX-M, 'PARallel TEXt-based language change measurement Method,' is provided in Section 3. In Section 4, we describe the stemming techniques we developed for Turkish and demonstrate their effectectiveness. Section 5 provides our experimental design with the description of the corpus. The experimental results on language change are given and discussed in Section 6. Section 7 concludes the article.

2 Related Works

Christiansen and Dale (2003) explain how some connectionist models can be used for computational modeling of language change. Juola (2003) presents an information theoretic model for measuring language change. He specifies no particular type of language change; however, he shows that meaningful measurements can be made from as few as 1000 characters. The use of words may also illustrate language change with time. For example, Woods (2001) shows that the most frequent word in modern Spanish was considerably less frequent during the sixteenth and seventeenth centuries.

A possible tool for language change studies is the use of objective literary style markers, such as the frequencies of most frequent words, and token and type length frequencies in text blocks. Based on such style markers statistical methods can be used to identify the characteristics of old and new texts or to distinguish them from each other. Such attributes are used in various authorship or stylometry studies (Baayen *et al.*, 1996; Binongo and Smith, 1999; Oakes, 1998). For example, Forsyth (1999) uses substrings for such purposes. In our recent stylometric studies (Can and Patton, 2004; Patton and Can, 2004) by using several style markers, including frequencies of most frequent words, and token and type lengths, we show that writing style changes in Turkish can be identified.

Another project, which is similar to our study, aims to describe and analyze the linguistic changes in old and modern French using the translations of works in classic Latin (Goyens and Van Hoecke, 1996).

Conceptually our approach (of employing old and new parallel translations and comparing them using statistical techniques to quantify the language change with time) is similar to the use of parallel texts, or bitexts, in language analysis. However, the bitext concept implies a source text and its translation in another language, but not in the same language. For example, Melamed's study (2001) shows how to obtain correspondence among tokens, sentences, passages, and how to determine translation omissions using bitext.

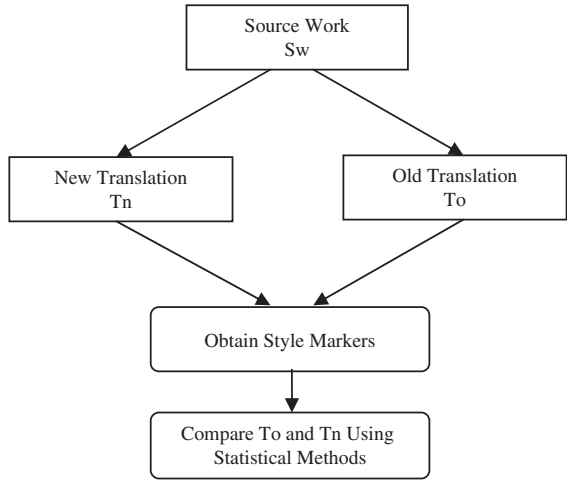

Fig. 1 Graphical description of PARTEX-M ('PARallel TEXt-based language change measurement Method')

3 PARTEX-M—PARallel TEXt-based Language Change Measurement Method

In PARTEXT-M, we use old and new parallel translations of foreign literary works in a certain target language whose change will be quantified. In PARTEXT-M, foreign works constitute the source. For each source work (Sw) we use old (To) and new (Tn) translations, and compare the unconsciously used language features of these translations (of a set of source works) using statistical methods. A graphical description of the method is provided in Fig. 1.

Our approach of using language features provides an objective comparison environment. These translations provide snap shots of the target language at different times. The aim of using translations is to eliminate the possible undesirable effects (such as the context and author bias) of works originally written in the target language. In a translation, what has to be written is well defined. However, there may be omissions and additions and changes of perceptions of a work's (or author's or genre's) significance. To overcome this we use multiple translated works printed by reliable publishers. The use of old and new parallel translations is an intuitive, efficient, and effective corpus sampling technique. Furthermore, works from different source languages filter unpredictable influences that can be introduced by a particular source language or work.

4 Turkish Language and Stemming for Turkish

As an application of PARTEXT-M, in this study we use the Turkish language. We first briefly introduce this language and then develop algorithms to obtain the stems to be used in the rest of the study. Stemmers and lemmatizers are two similar, but different language tools. A lemmatizer tries to find the dictionary entry of a word; in contrast, a stemmer obtains the root in which a word is based. Due to the nature of English, sometimes words are mapped to lemmas which apparently do not have any surface connection as in the case of *worse* and *worst* being mapped to *bad*. However, Turkish does not have such irregularities and it is always possible to find the 'stem' or 'lemma' of any given word through application of grammar rules in removing the suffixes. For this reason, throughout the article, we prefer the word 'stemming' over lemmatization; as it is more commonly used, and our algorithms internally identify the suffixes and remove them in the stemming process.

4.1 Turkish language

Turkish is an agglutinative language similar to Finnish. 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 word suffixes.

In contemporary everyday Turkish, it is observed that words have about three to four 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 thirty-three. 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 are 32, 478, and 4,789. For 'verb' type word stems the numbers are 46, 895, and 11,313 (Hakkani-Tür, 2000, p. 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) and Solak and Oflazer (1993). A two-level (lexical and surface) morphological description of Turkish word structure is studied in (Oflazer, 1994). Statistical modeling and its use in morphological disambiguation, spelling correction, and speech recognition are studied in (Hakkani-Tür, 2000).

4.2 Stemming for Turkish

Several researchers have worked on stemming in Turkish (Solak and Can, 1994; Alpkoçak *et al.*, 1995; Duran, 1997; Ekmekçioğlu and Willett, 2000). Turkish stemming methods usually return more than one result and do not select the best stem among the possible candidates for a given word. Although it does not directly address stemming, Oflazer's morphological analyzer (1994) gives all possible analyses for a given word based on a stem list and structural analysis. A recent study by Hakkani-Tür (2000) reports on statistical methods for disambiguation of Turkish. However, disambiguation is a more complex task that includes much deeper analysis that may be unnecessary in stemming. In this study, we basically aim to find the correct stem among all possible alternatives. In order to select the best stem, we introduce two approaches (Altintas and Can, 2002).

4.2.1 *Stemming based on disambiguated corpus stem length information*

In this approach, we investigate four different stemming methods by using the average stem length information obtained from a disambiguated corpus supplied by Bilkent University (TLSPC, 2004). It will be referred to as the 'Bilkent corpus'. We also have an additional, the fifth, method which does not pay attention to the average stem length information.

The total number of tokens in the Bilkent corpus is 712,272. The number of types is 108,875, and distinct number of stems for types is 24,388. First 250, most frequent distinct stems constitute 47% of the corpus. Average stem length of tokens and types, respectively, are 4.58 and 6.58 characters. More than half of the words are nouns and one-fifth are verbs. Table 1 provides the frequency of appearance of each part of speech (POS) in the corpus.

Both the Bilkent corpus and the test data (defined in the next section) were analyzed by using Oflazer's morphological analyzer (Oflazer, 1994). In the results of the analyzer, the first morpheme is the root of the corresponding analysis followed by POS information. Then other morphemes come to form the analysis.

In this part, we analyzed the data morphologically. All possible analyses were sent to the

Table 1 Frequency and % occurrence for each part of speech (POS) in the Bilkent corpus

Part of speech	Frequency	% Occurrence
Nouns	388,665	54.567
Verbs	142,618	20.023
Adjectives	56,658	7.955
Conjunctives	34,677	4.867
Determiners	23,620	3.316
Adverbs	20,297	2.850
Post positions	15,997	2.246
Pronouns	14,880	2.089
Numbers	12,410	1.742
Questions	1,898	0.266
Interjections	430	0.060
Duplications	122	0.017

appropriate functions, representing each method we used for stemming. We used five different methods.

- Returning the stem of the analysis that is returned first by the morphological analyzer as the result. There is no specific ordering of the morphological analyses [personal communication with Kemal Oflazer]. (1: *First Found Method* or *Any Length Method*)
- Comparing the lengths of the stems of the possible analyses with the average stem length for tokens (4.58) and average stem length for types (6.58) and choosing the stem with the closest length to the average. (2: *Avg. Token Method*, 3: *Avg. Type Method*)
- Whenever there is more than one result with the same length, the part of speech information of the stem is considered, and the stems are given precedence according to their POS information in the order given in Table 1. (4: *Avg. Token with POS Info. Method*, 5: *Avg. Type Stem with POS Info. Method*)

Table 2 summarizes the experimental results. The test data is approximately 20,000 words randomly selected from the unambiguous Bilkent corpus. The test data was not included in the training set. The correct answers are those that have the same root and POS with what is reported in the corpus. The second column of Table 2 provides the number (success rate) of each stemming algorithm. The third column provides the same information with the correct stem disregarding the POS. Table 2 shows that the methods produce similar results. Having a result of around 90% may be imperfect, but could be acceptable for many applications. The length-based method is simple to implement provided that there is a morphological analyzer available.

Table 2 Results for stem length-based stemming methods

Method	Stem and POS correct	Stem correct and POS ignored
First found (Any length)	15,506 (76.2%)	16,677 (81.9%)
Avg. token stem	15,870 (77.9%)	17,919 (88.0%)
Avg. type stem	16,398 (80.5%)	18,468 (90.7%)
Avg. token with POS info.	16,552 (81.3%)	17,972 (88.3%)
Avg. type with POS info.	17,099 (84.0%)	18,520 (91.0%)

4.2.2 *Statistical stemming based on the n-gram language models*

In the statistical stemming part, we used the unigram, bi-gram and tri-gram language models (Ney *et al.*, 1994). The unigram language model calculates the probability of a word based on its frequency in a given corpus, regardless of the context information. The bi-gram language model tries to approximate the probability of a word, given all of the previous words, by the conditional probability of the preceding word. In general, the *n*-gram language model tries to approximate the probability of a word based on the conditional probability of the previous $(n-1)$ words.

For the statistical part of the experiment, the amount of data necessary to conduct the research is much larger than the stem length-based approach. The training data was extracted using the corpus available from Tür and Hakkani-Tür (Personal communication, 2002). The corpus was collected from Milliyet Newspaper covering the period from 1 January 1997 through 12 September 1998. There are around 20 million tokens in the 'Milliyet corpus' and the number of words, excluding sentence boundary tags and other unnecessary information, is about 18 million. We trained the system for words with and without part of speech information. The tokens were again analyzed by Oflazer's system (1994).

Tokens with a single alternative are used as they are, and ambiguous tokens are changed to the token <AMB>. For example, the word '*gülüm*' (my rose/I am a rose) has two morphological analyses both of which are derived from the root '*gül+Noun*' (rose+Noun). So, this word is tokenized as 'gül+Noun' when POS information is considered. However, the word 'güldür' (S/he/it is a rose/Cause them to smile) has also two analyses, which are derived from two distinct roots '*gül+Noun*' (rose+Noun) and '*gül+Verb*' (smile+Verb). Thus, this word is changed to the token <AMB> when POS is considered and is saved as *gül* when POS is not considered. The number of tokens and *n*-grams can be seen in Table 3.

We used two texts for testing purposes. In order to prevent any possible bias, we refrained from using the text of the language change experiments

Table 3 The number of tokens and *n*-grams in the *Milliyet* Corpus

	No. of tokens excluding unnecesary tags	No. of ambiguous tokens	Unigrams	Bi-grams	Tri-grams
With POS info.	~18 M	5,411,084	89,764	1,490,322	1,456,709
Without POS info.	~18 M	2,374,760	50,200	1,217,744	1,136,253

and instead used two independent texts: (i) a passage from Yaşar Kemal's '*İnce Memed* (Vol. 1)' (IM1) with 4,268 tokens, and (ii) a collection of some newspaper articles from the year 2002 with 1,872 tokens. Words in both texts were tagged manually by a human expert for their roots and are assumed 100% correct. In the experiments, we used the SRI Language Modeling Toolkit for statistical processing (SRI, 2004).

Table 4 provides the results. Its last three columns show the percentage of the correct stems with different methods. The table shows that results without POS information are better than those with POS information. This is because many words have the same root with different POS. For example, the word '*bir*' (one) has four analyses all of which have the same root: *bir+Adv*, *bir+Adj*, *bir+Num+Card*, *bir+Det*.

The results for the newspaper articles are slightly better than that of IM1. This is probably due to the training data, which is collected from a newspaper. In general, the domain of the corpus directly affects the results (Jurafsky and Martin, 2000, p. 202). For example, IM1 includes many proper names, which are valid Turkish words, but are not recognized by the morphological analyzer. However, note that the performance difference of the methods with the IM1 and the newspaper articles is insignificant. This intuitively implies that the methods can confidently be used with other types of text.

Many of the wrongly recognized words appear in the stop word list for Turkish by Tür (Tür, 2000, Appendix B). For example, words such as *önce* (before), *üzerine* (after having done so), *için* (for), *ile* (with) are accepted to be stop words. All of these words have more than one analysis and thus are tagged as <AMB> in the corpus and do not count towards the disambiguation. If the stemming is used for information retrieval, such words should be excluded and the system performance may increase considerably.

Table 4 Results for statistical stemming

	No. of tokens	Correct results with unigram	Correct results with bi-gram	Correct results with tri-gram
IM1 with POS	4268	86.4%	86.7%	86.5%
IM1 without (w/o) POS	4268	92.2%	92.4%	92.3%
Newspaper articles with POS	1872	87.2%	88.0%	88.1%
Newspaper articles w/o POS	1872	91.4%	92.5%	92.4%

We have not used any preprocessing for the training data, all words were processed as they appear in the corpus. A preprocessor can be used to eliminate some of the ambiguous analyses. This can improve the system performance.

Table 4 shows that tri-gram results are not better than bi-gram results. Table 3 shows that the number of tri-grams for both experiments is less than that of bi-grams. This is due to both ambiguities in the training data and the data sparseness. If we had more training data that would allow us to construct a larger number of tri-grams, we could expect better results for the tri-gram case. In the language change experiments, we use the bi-gram stemming approach without using the POS information. Our unigram and tri-gram approaches can also be used for the same purpose; they provide almost the same level of stemming effectiveness as the bi-gram approach as shown in Table 4.

5 Experimental Environment and Design

The previous section describes the process of obtaining stems. From this, we can obtain stem lengths and suffix lengths. These and other style

markers are necessary components of PARTEX-M. Our source languages are English, French, and Russian. The source works are also of different varieties including essays, novels, and plays. We aim for diversity in our corpus to achieve better representation of the target language usage. Appendix Table A1 shows the details of the translations. It includes the acronyms, such as BG-1957, corresponding to the translations. The old and new translations all together provide a total text size of 244,510 tokens. For our discriminant and logistic regression analyses, both defined later, we decided to subdivide each work into 1,000 word blocks as units in our statistical experiments. This block size is large enough for our analyses, yet small enough to provide, at least nine blocks from each work (Binongo and Smith, 1999, p. 460; Forsyth and Holmes, 1996, p.164; Baayen *et al.*, 1996, p.122). At the same time, the use of blocks rather than complete works gives the opportunity to examine the works at a micro-level. The use of complete works in our analysis allows us to conduct additional experiments at the macro-level.

Our aim is to examine the change in the quantifiable features of a language. In this particular case, our focus is Turkish. We designed the experiments for both tokens and types. Doing the experiments only for tokens may not give complete information, because repetitions in the corpus might cause a wrong interpretation of the results. Furthermore, using only the surface forms of words may be insufficient, because Turkish is an agglutinative language, and meaning is enriched by concatenation of suffixes to a stem. So, we performed the experiments both for the surface and stemmed forms of the tokens and types. All of these analyses were conducted using the SAS for Windows software, Version 9.

6 Experimental Results

6.1 Changes related to number of tokens, types, and vocabulary richness

Table 5 provides the results of the measurements for surface forms. A matched paired *t*-test was conducted to determine differences in the number of tokens between the old and new translations of each work for both surface forms and stem forms. Using a significance level of 0.05 the test concluded that there is no significant difference. Therefore, we cannot make a generalization for the change in number of tokens.

Table 5 Results for surface forms*

Work acronym	No. of types	No. of tokens	Type to token ratio
BG-1957	4,966	12,511	39.69
BG-1999	5,305	13,845	38.32
D-1947	4,607	9,907	46.50
D-2002	4,617	9,609	48.05
DM-1944	13,065	36,398	35.90
DM-1990	12,077	33,007	36.59
H-1944	9,411	25,668	36.66
H-1999	8,571	25,121	34.12
M-1946	5,946	14,754	40.30
M-1999	5,630	14,352	39.23
UK-1954	4,223	11,911	35.46
UK-1999	5,062	12,843	39.42
YK-1943	5,146	12,526	41.08
YK-1999	4,587	12,058	38.04

*Adjacent pairs with the same prefix (e.g. BG, D, etc.) are old and new translations of the same work (see Appendix Table A1 for more information).

Table 6 [21] shows the change of the same language features in terms of stems. It shows that the number of types has decreased considerably for all cases. We think that the vocabulary of the language has shrunk over time, and today we have fewer root words than we had in the past.

For measuring the change in terms of vocabulary richness of the old and new translations, we use the TTR, i.e. (no. of types)/(no. of tokens) in a given translation. We multiply this ratio by 100 to express it as a percentage change (we still call it TTR). The TTR has been criticized in the literature, because the ratios obtained are variable and related to the number of tokens in the sample text (McKee *et al.*, 2000; Tweedie, Baayen, 1998). However, notice that in our case, paired old and new translations are based on the same source text and we found no significant difference in the number of tokens between the old and new translations. Thus, it makes sense to use the TTR as a measure to quantify the language change between old and new translations. We use TTR at two different

Table 6 Results for stems*

Work acronym	No. of types	No. of tokens	Type to token ratio
BG-1957	1,914	12,508	15.30
BG-1999	1,631	13,843	11.78
D-1947	1,634	9,905	16.50
D-2002	1,537	9,605	16.00
DM-1944	4,983	36,382	13.70
DM-1990	3,857	32,995	11.69
H-1944	3,709	25,656	14.46
H-1999	2,728	25,109	10.87
M-1946	2,067	14,744	14.02
M-1999	1,704	14,342	11.88
UK-1954	1,529	11,908	12.84
UK-1999	1,490	12,838	11.61
YK-1943	2,160	12,523	17.25
YK-1999	1,661	12,058	13.78

*Please see endnote no. 1 (at the end before Appendix).

levels: (i) for the surface level tokens and types without stemming (surface-TTR), (ii) for the stemmed tokens and types (stem-TTR). The surface-TTR in general shows a decrease as we go from old to new translations (for the works: BG, H, M, and YK). However, the stem-TTR shows a decrease for all cases. The average stem-TTRs for the old and new translations were 14.867 and 12.516, respectively. A one-way analysis of variance was conducted to detect whether these average stem-TTRs are significantly different. Using a significance level of 0.05, the test concluded this difference to be strongly significant with an observed significance level (*P*-value) of 0.02.

6.2 Changes related to token and type lengths

6.2.1 *Discriminant analysis*

To provide further motivation to our later hypothesis tests, a series of discriminant analyses were conducted on the translations of each of the seven works to determine how well token word lengths could discriminate the old from the new translations. Blocks of 1,000 words made up each experimental unit. Frequencies of token lengths from 1 to 20 characters served as potential discriminators. A stepwise discriminant analysis was conducted to determine what token length frequencies provide the best separation between the work types.

The average correct classification rate over all of the analyses was 80%. This was calculated by dividing the total number of successful classifications by the total number of old and new blocks over all seven works. This indicates that language change has taken place from the period between the old and new translations relative to the style markers, token, and type lengths.

6.2.2 *Logistic regression analysis*

The classification of the translation is treated as a binary variable (old, new). To determine whether significant differences in the frequencies of the token and type length existed between the two classification types, a series of logistic regressions were conducted using the classification of the translation as the dependent variable and the frequencies of the token or type lengths as the independent variable for a given block. The regressions were done separately for tokens and types. We restricted our experimental region of token and type lengths to no more than seventeen characters since longer words were very sparse in the corpus, and in general, in Turkish (Dalkılıç and Çebi, 2003).

The results of these logistic regressions are given in Appendix Table A2. Appendix Table A2 contains data for the non-Shakespearean (Panel A) and Shakespearean work (Panel B). For each of the seven works, the average number of occurrences of token and type lengths per block is given in separate columns. The columns adjacent to these contain the odds ratio output from the logistic regression. The odds ratio is a measure of association and compares the odds of finding a word belonging to an old translation to the odds of belonging to a new translation when that word, having a stem of a certain length, is chosen at random. An odds ratio less than one indicates that such a word is more likely to come from an old translation, whereas a ratio greater than one indicates a greater likelihood that it is from a new one. The large number of hypothesis tests conducted by the logistic regressions lead to problems with alpha significance levels. To reduce the number of tests, we conducted

Table 7 Regression results for token lengths

Author	Regression equation	*F*-value	*P*-Value	R^2
Daudet	Log(odds ratio)=−0.029+0.005×token length	$F(1, 10)=5.91$	0.0354	0.371
Dostoyevsky	Log (odds ratio)=−0.123+0.022×token length	$F(1, 22)=15.14$	0.0008	0.408
Montaigne	Log(odds ratio)=−0.104+0.018×token length	$F(1, 10)=11.19$	0.0074	0.528
Shakespeare	Log(odds ratio)=−0.033−0.007×token length	$F(1, 34)=3.11$	0.0867	0.084

Table 8 Regression results for type lengths

Author	Regression equation	*F*-value	*P*-Value	R^2
Daudet	Log(odds ratio)=−0.152+0.019×type length	$F(1, 10)=7.43$	0.0213	0.426
Dostoyevsky	Log (odds ratio)=−0.167+0.031×type length	$F(1, 22)=9.84$	0.0048	0.309
Montaigne	Log(odds ratio)=−0.052+0.013×type length	$F(1, 10)=7.35$	0.0219	0.424
Shakespeare	Log(odds ratio)=−0.030−0.001×type length	$F(1, 34)=0.01$	0.9077	0.0004

separate ordinary least squares (OLS) regressions on the tokens data and the types data using the natural log of the odds ratio as the response variable. The natural log transformation applied to the odds ratio converts a non-negative variable to the one that has a more expanded range encompassing both positive and negative values. (The idea for this type of regression came from a suggestion made by an anonymous referee of (Can, Patton, 2004).) Both word length and author were the independent variables. We also included an interaction term between author and word length. In general, an interaction between two factors, A and B, indicates that the effect of Factor A is dependent on the level of Factor B. In two of Shakespeare's works (*Hamlet*—H and *Comedy of Errors*—YK), the average token and type word length are both less in the new translation than in the old. Since the opposite is true with the other authors, we felt there was a need to test for an interaction effect. Types and tokens containing more than twelve characters were excluded due to their small number (especially in Shakespeare's works).

An initial analysis of variance performed on the token data indicated a very significant word length effect [$F(1, 83)=11.03$, $P=0.0014$]; a very significant author effect [$F(3, 83)=4.51$, $P=0.0058$], and an extremely significant interaction effect [$F(3, 83)=8.70$, $P<0.0001$].

Since the interaction effect had extremely strong significance, individual simple regressions were conducted for each author using token length as the independent variable. Table 7 summarizes the results.

With the exception of Shakespeare, the regression analysis for each author had significant token length effects. Since the coefficient estimates to token length in these regressions were positive, a longer token would have a higher probability of belonging to a new translation.

A similar analysis was conducted on the type data. We got strong significant results that were perhaps not as dramatic as the token results. Again, a preliminary analysis of variance was performed on the type data. The results indicated a very significant type length effect [$F(1,83)=10.59$, $P=0.0017$]; an insignificant author effect [$F(3,83)=1.33$, $P=0.2707$], but a significant interaction effect [$F(3,83)=0.0292$, $P=0.0292$]. Due to the strong significance of the interaction effect, individual simple regressions (again using type length as the independent variable) were conducted for each author. Table 8 summarizes the results.

Based on the type data, the regression analysis for each author (except Shakespeare) had significant type length effects. Since the coefficient estimates to type length in these regressions were positive, a longer type would have a higher probability of belonging to a new translation.

From the regression equations in Tables 7 and 8, we can get the predicted odds ratio as a function of token and type length for each author.

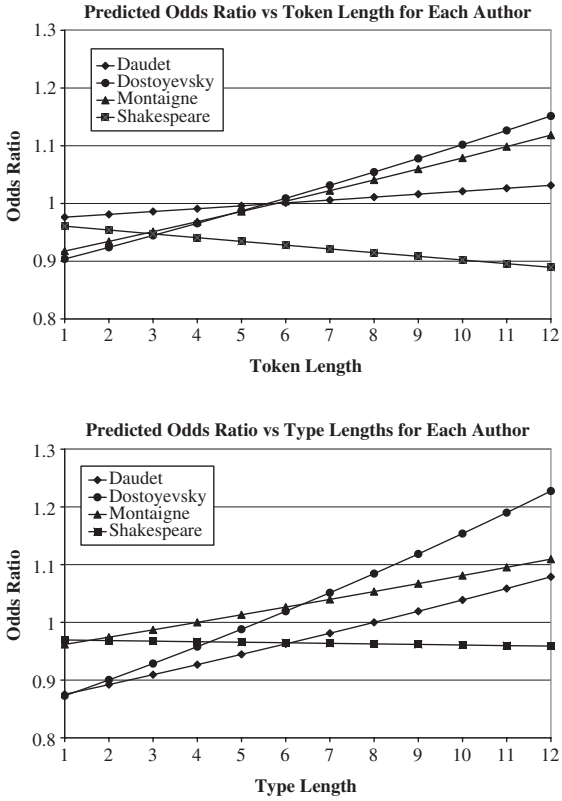

Fig. 2 Predicted odds ratio for token and type lengths for each author

As an example, the prediction odds ratio as a function of token length for Daudet would be the following.

$$\text{Predicted odds ratio} = e^{-0.029+0.005\times\text{token length}}$$

A series of graphs showing the predicted odds ratio plotted for each author against token and type lengths appear in Fig. 2. In interpreting these graphs, assume that a word is chosen at random from a block of one of the translations for a given author's work. If the predicted odds ratio for that token length is greater than one, the chances are greater that the block itself comes from a new translation rather than an old one. Likewise if a vocabulary word, i.e. type, is chosen at random from a block of a translation for a given author's work, the same interpretation applies. With the exception of Shakespeare, the predicted odds ratio for both tokens and type increase as the length increases.

6.3 Changes related to suffix and stem lengths

6.3.1 *Changes related to suffix lengths*

Table 9 provides information for token and type average suffix lengths, average stem lengths, and average word lengths. Using the data from this table, a one-way analysis of variance was conducted to determine whether there is change in the suffix type lengths between the old and newer translations. A significance level of 0.05 was used. The average type suffix lengths of the old and new translations were 2.026 and 2.509, respectively. The observed significance level of this difference was 0.046 indicating strong evidence of longer type suffix lengths in the newer translations. A similar analysis was conducted for tokens. The average suffix lengths of tokens for both old and new were 1.933 and 2.104, respectively. However, this difference was not statistically significant since the observed significance level was greater than 0.05

6.3.2 *Changes related to stem lengths*

Table 9 shows that as we go from old translations to new, for a given work, both the token and type stems become shorter. This is interesting because as we go from old to new translations the average token and type lengths tend to increase. This together with the decrease in the number of stems shows us that the vocabulary of the language has changed considerably with time. In newer words, on the average, stems are shorter and suffixes are longer. This means that more meaning has been loaded into a single stem by using more number of suffixes for that stem.

To study the nature of the change, a series of logistic regressions were conducted where the binary response variable for each was the classification of the translation (old, new). The independent variable was the frequency of tokens or types of a certain stem length for a given block. The results of these logistic regressions are given in Appendix Table A3. Panel A of Appendix Table A3 contains the data for the works of the authors other than Shakespeare and Panel B

Table 9 Averages of token and type lengths, and their stem and suffix lengths

Work acronym	Avg. token length (atol)	Avg. token stem length (atosl)	Avg. token suffix length (atol−atosl)	Avg. type length (atyl)	Avg. type stem length (atysl)	Avg. type suffix length (atyl−atysl)
BG-1957	5.96	3.95	2.01	7.85	5.82	2.03
BG-1999	6.04	3.78	2.26	8.01	5.39	2.62
D-1947	6.20	3.88	2.32	8.00	5.31	2.69
D-2002	6.32	3.82	2.50	8.08	5.16	2.92
DM-1944	6.01	4.19	1.82	7.88	6.32	1.56
DM-1990	6.07	4.09	1.98	7.97	5.80	2.17
H-1944	5.96	4.24	1.72	7.85	6.53	1.32
H-1999	5.73	3.92	1.81	7.72	5.55	2.17
M-1946	5.84	3.95	1.89	7.60	5.31	2.29
M-1999	5.91	3.88	2.03	7.71	5.07	2.64
UK-1954	5.83	3.86	1.97	7.84	5.44	2.40
UK-1999	6.11	3.76	2.35	8.01	5.20	2.81
YK-1943	5.88	4.08	1.80	7.60	5.71	1.89
YK-1999	5.62	3.82	1.80	7.31	5.08	2.23

corresponds to the Shakespearean works. These Panels of Table A3 have a similar structure as that of Appendix Table A2; the difference is that Panels A & B of Appendix Table A2 refer to word lengths whereas Panels A & B of Appendix Table A3 refer to stem lengths. Words having stem lengths up to twelve characters were used since words having longer stems were very sparse in the corpus. The natural log of the odds ratios was used as a dependent variable in OLS regressions that had author and stem length as independent variables. One regression was done for the token data and another for types. This type of analysis was not attempted on suffix lengths due to its limited range of values.

Besides an interest in stem length effects on the odds ratio, we were also interested in the author effect and its interaction with stem length. As shown in Appendix Table A1, some individuals translated more than one work. However, we neglected the translator effect in this analysis since most of the translators handled only one work.

In this analysis, we used stem lengths up to eight characters since longer stem lengths had very small average occurences (less than ten per block in most works, see panels A & B of Appendix Table A3). A preliminary analysis found neither a significant author effect nor an interaction effect but did find a significant stem length effect. This was true for both the token and type data. Thus, we developed our models based on stem length alone as the independent variable. Upon inspecting the residuals and the odds ratio in Appendix Table A3, we observed the odds ratio had a tendency to increase for stems of length one to four and then decrease for stems of lengths greater than four. We subsequently developed a quadratic regression model with linear and quadratic stem length terms as independent variables and the natural log of the odds ratio as the dependent variable.

Our regression results for both tokens and type stems indicated an extremely strong relationship between stem length and log of the odds ratio. The tokens regression produced an overall $F(2,53)=22.99$ $(P<0.0001)$. The prediction equation for the token's regression was the following.

$$\begin{aligned}\text{Log(odds ratio)} = &-0.129 + 0.08928 \\ &\times \text{stem length} - 0.1208 \\ &\times (\text{stem length})^2\end{aligned}$$

The linear and quadratic regression coefficient estimates both had observed significance levels of $P<0.0001$ indicating extremely strong evidence of a positive linear stem length coefficient and a negative quadratric coefficient. Analyzing the prediction equation, the log of the odds ratio tends to increase as the stem length increases from one to four, and then decreases to negative values for increases beyond 4. Hence, tokens having longer stem lengths

have a higher probability of belonging to the old translation. Since there was not a significant interaction effect between stem length and author, this property appears to be uniform across all of the four authors. The coefficient of determination (R^2) statistic was 0.4645 indicating that 46.45% of the total variance of the odds ratio log about its mean can be explained by token stem length. There definitely are other factors besides stem length affecting the odds ratio, but stem length is a very important factor.

We obtained similar results for the type data (i.e. types having longer stem lengths have a higher probability of belonging to the old translation). The types regression produced an overall $F(2,53)=13.27$ ($P<0.0001$). The prediction equation for the type's regression was the following.

$$\begin{aligned}\text{Log(odds ratio)} = &-0.850 + 0.4353 \\ &\times \text{stem length} - 0.05031 \\ &\times (\text{stem length})^2\end{aligned}$$

Both the linear and quadratic estimates yielded observed significant levels of $P<0.0001$. The R^2 statistic was 0.3336, which was not quite as strong as the token case but strong nevertheless.

The predicted odds ratio as a function of token and type stem lengths can be obtained by exponentiating both sides of each regression equation. Figure 3 contains the plots of the odds ratio against stem length for both token and type stems. In both of these, the predicted odds ratio is largest for stems of approximately length four. For stems greater than four, the odds that a block is selected from a new translation decreases as stem length increases. Stems having lengths of three, four, or five have a greater chance of coming from new translations. It is interesting to note that very short stems, having lengths one or two tend to appear in older translations. However, the average occurrences of these stems are relatively small (Appendix Tables A3).

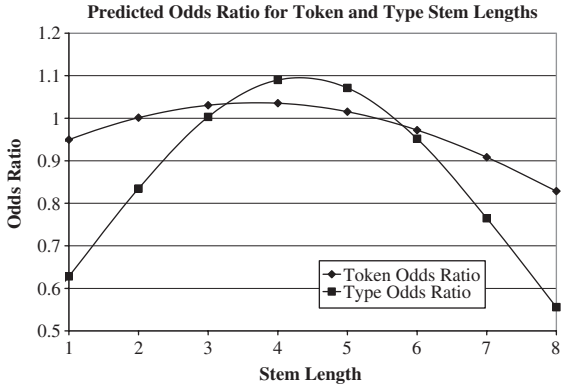

Fig. 3 Predicted odds ratio for token and type stem lengths

7 Conclusions

In this study, we introduce various stemming techniques for Turkish and a systematic method, PARTEXT-M (PARallel TEXt-based language change measurement Method), for quantifying language change. In agglutinative languages like Turkish, stemming is important in measuring language change in terms of some style markers, since a single word stem may yield many different surface forms. Our approach to stemming in Turkish can be applied to some other agglutinative languages. The successful results with Turkish indicate that PARTEX-M is promising for quantifying change in other languages.

The experiments show that there is a decrease in vocabulary richness when measured as TTR using word stems. Hypothesis tests indicate a strong significant increase in the suffix lengths of types going from the older to the newer translations. For newer translations, stem lengths tend to be shorter and types and token lengths tend to be longer. Since the number of tokens of the old and new translations is not significantly different, these observations indicate that in contemporary Turkish one would use more suffixes to compensate for the fewer stems to preserve the expressive power of the language at the same level. This is in harmony with our vocabulary richness (stem TTR) result that indicates a decrease in going from old to new. The increase in suffix lengths and decrease in stem level vocabulary richness can be partly explained by neologisms introduced for replacing old words in contemporary Turkish. Such neologisms are usually obtained by adding suffixes to Turkish stems

(i.e. by only using stems which are not borrowed from other languages).

The PARTEX-M approach uses time-separated parallel translations to quantify diachronic change in a target language. Frawley (1984) considers translations as 'third code', a code which is different from both source and target language. [Here one may also recall the phrase '*Traduttore, traditore*'('the translator is a betrayer') (Jakobson, 1959).] Based on the 'third code' concept, one can claim that 'a translation is at best an unrepresentative variant of the target language. As such, it is misleading to generalize the results based on such biased data to the target language. The effects of translation process on the translated text are unavoidable.' By following this line of thinking, users of PARTEX-M should be careful for potential problems. Whilst, Even-Zohar (1990) regards translated literature as a system of its own, in view of the fact that we have multiple parallel translations, it is fair to say that the changes in the translations are 'at least' the reflections of the changes in the target language (Turkish). Since the sources are the same, the changes in the translations should or can be attributed to the changes of the target language. Of course, a balanced diachronic corpus that covers a wide range of genres and a large number of authors can certainly minimize such criticism or possible problems. However, such an approach involves two major undertakings: creation of this diachronic corpus, and repetition of our experiments by using this new corpus. This is an interesting future research possibility. The study reported by Tirkkonen-Condit (2002) illustrates that in Finnish the translations can be 'not readily distinguishable' from originally produced (non-translated) text. The identicalness of translated (translational data) and non-translated (original) texts can be investigated in Turkish. The study of the 'third code' concept (Overas, 1998) in Turkish translations is another interesting challenge for researchers.

Acknowledgements

We appreciate the anonymous referee comments that brought the concept of 'third code' and some other issues to our attention. We would also like to thank Varol Akman, Tuncay Birkan, Kemal Oflazer, Gökhan Tür, Dilek Zeynep Hakkani-Tür, Pedrito Uriah Maynard-Zhang, and Bilkent University.

References

Aitchison, J. (2001). *Language Change: Progress or Decay?* 3rd edn. Cambridge and New York: Cambridge University Press.

Alpkoçak, A., Kut, A. and Özkarahan, E. (1995). Bilgi bulma sistemleri için otomatik Türkçe dizinleme yöntemi. In the *Proceedings of 12. Ulusal Bilişim Kurultay*ı, 247–53.

Altintas, K. and Can, F. (2002). Stemming for Turkish: a comparative evaluation. In the *Proceedings of the 11th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN)*, (İstanbul, Turkey, June 2002), Istanbul: Istanbul University Press, pp. 181–8.

Baayen, H., Halteren, H. V. and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**(3): 121–32.

Binongo, J. N. and Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, **14**(4): 445–66.

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

Carroll, T. (2001). *Language Planning and Language Change in Japan*. Great Britain: Curzon Press.

Christiansen, M. H. and Dale, R. (2003). Language evolution and change. In Arbib, M. A. (ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press, pp. 604–6.

Dalkılıç, G. and and Çebi, Y. (2003). Creating a Turkish corpus and determining word length distributions that are used in Turkish texts. *Dokuz Eylül University Science and Engineering Journal*, **5**(1): 1–7.

Duran, G. (1997). *Gövdebul: Turkish Stemming Algorithm*. M.S. Thesis, Department of Computer Engineering, Hacettepe University, Ankara, Turkey.

Ekmekçioglu, Ç. and Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. *Program*, **34**(2): 195–200.

Even-Zohar, I. (1990). The position of translated literature within the literary polysystem. *Poetics Today*, **11**: 45–51.

Frawley, W. (1984). Prolegomenon to a theory of translation. In William, Frawley (ed.), *Translation: Literary, Linguistic, and Philosophical Perspectives*. Newark: University of Delaware Press, pp. 159–175.

Forsyth, R. S. (1999). Stylochronometry with substrings, or: a poet is young and old. *Literary and Linguistic Computing*, **14**(4): 467–78.

Forsyth, R. S. and Holmes, D. I. (1996). Feature finding for text clarification. *Literary and Linguistic Computing*, **14**(4): 168–173.

Goyens, M. and Van Hoecke, W. (1996). Traduction et changement linguistique: une étude empirique de l'évolution des possessifs du latin au français moderne. *Bien Dire et Bien Aprandre*, **13**: 39–58.

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

Holt, D. E. (ed.) (2003). *Optimality Theory and Language Change*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, **37**(1): 77–96.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing*. Prentice Hall.

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

McKee, G., Malvern, D. and Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, **15**(3): 323–38.

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

Melamed, D. (2001). *Empirical Methods for Exploiting Parallel Text*. Cambridge, MA: MIT Press.

Ney, H., Essen, U. and Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modeling. *Computer Speech and Language*, **8**(1): 1–38.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, **9**(2): 137–48.

Oflazer, K. (2003). Dependency parsing with an extended finite-state approach. *Computational Linguistics*, **29**(4): 515–44.

Overas, L. (1998). In search of the third code: an investigation of norms in literary translation. *Meta*, 43 (4):571-588. http://www.erudit.org/revue/meta/1998/v43/n4/003775ar.html (accessed 20 April 2005).

Patton, J. M. and Can, F. (2004). A stylometric analysis of Yaşar Kemal's *İnce Memed* tetralogy. *Computers and the Humanities*, **38**(4): 457–67.

Solak, A. and Can, F. (1994). Effects of stemming on Turkish text retrieval. *In Proceedings of the 9th Int. Symp. on Computer and Information Sciences*. Turkey: Antalya, pp. 49–56.

Solak, A. and Oflazer, K. (1993). Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing*, **8**(3): 113–30.

SRI. (2004). SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/ (accessed 18 June 2004).

Tirkkonen-Condit, S. (2002). Translationese – a myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target*, **14**(2): 207–20.

Tür, G. (2000). *A Statistical Information Extraction System for Turkish*. Bilkent University, Department of Computer Engineering, Ph.D. Thesis, Ankara, Turkey.

Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, **32**(5): 323–52.

TLSPC. (2004). Turkish Language and Speech Processing Center. http://www.nlp.cs.bilkent.edu.tr (accessed 18 June 2004).

Woods, M. J. (2001). Spanish word frequency: a historical surprise. *Computers and the Humanities*, **35**(2): 231–6.

Note

1. In Tables 5 and 6 the number of types and number of tokens for some corresponding entries are not exactly the same. Although the difference is negligibly small, it deserves an explanation. While finding the number of types and number of tokens, we omit Arabic numerals. During the morphological analysis, the Oflazer's system converts Roman numerals into Arabic numerals. Consequently, some numbers are counted in surface forms but they are not counted in stems.

Appendix

The translations used in the experiments are provided in Appendix Table A1. After each author (e.g. Daudet) we provide: the Turkish title of the work (*Değirmenimden Mektuplar*), its English title (Letters from my Windmill) in parentheses—if needed, after that for each translation, we provide its acronym (such as DM-1944), the name of the translator (such as Sabri Esat Sivayuşgil), the publisher of the translation, the publication place and year.

Table A1 The source works used in the study

Alphonse Daudet
- *Değirmenimden Mektuplar* (Letters from my Windmill)
  - DM-1944: Sabri Esat Sivayuşgil, Milli Eğitim, Ankara, 1989.[1]
  - DM-1990: Rabia Ergüven, İnkilap Kitabevi, İstanbul, 1990.

Fyodor Dostoyevsky
- *Beyaz Geceler* (White Nights)
  - BG-1957: Nihal Yalaza Taluy, Varlık Yayınları, İstanbul, 1957.
  - BG-1999: Mehmet Özgül, Cumhuriyet Gazetesi, İstanbul, 1999.
- *Uysal Kız* (The Gentle Maide)
  - UK-1954: D. Sorakın, S. Aytekin, Maarif, Ankara, 1954.
  - UK-1999: Mehmet Özgül, Cumhuriyet Gazetesi, İstanbul, 1999.

Michel de Montaigne
- *Denemeler* (Essays)[2]
  - D-1947: Sabahattin Eyüboğlu, Milli Eğitim, Ankara, 1947.
  - D-2002: Celal Öner, Oda Yayınları., İstanbul, 2002.

William Shakespeare
- *Hamlet*
  - H-1944: Orhan Burian, Maarif, Ankara, 1944.
  - H-1999: Bülent Bozkurt, Remzi Kitapevi, İstanbul, 1999.
- *Macbeth*
  - M-1946: Orhan Burian, Milli Egitim, Ankara, 1946.
  - M-1999: Orhan Burian (Edited by Publisher), Cumhuriyet Gazetesi, İstanbul, 1999.
- *Yanlışlıklar Komedyası* (Comedy of Errors)
  - YK-1943: Avni Givda, Maarif, Ankara, 1943.
  - YK-1999: Bülent Bozkurt, Remzi Kitapevi, İstanbul, 1999.

Notes:
[1]The 1989 edition of Siyavusgil's translation is identical with his translation that was published in 1944 and the acronym we use for this work is DM-1944.
[2]We only use the common essays of D-1947 and D-2002.

Table A2 Logistic regression results comparing token and type lengths between old and new translations

Panel A: For the works of Daudet, Dostoyevsky, and Montaigne

		B G				D				D M				U K			
		Tokens		Types		Tokens		Types		Tokens		Types		Tokens		Types	
Word Length	Work type	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds Ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio
1	Old	35.00	0.941	3.83	0.943	14.67	0.905	3.89	0.937	23.94	0.964	5.50	0.715	32.18	0.879	5.09	0.54
	New	19.69		3.69		10.11		3.78		20.61		4.85		23.25		4.50	
2	Old	60.67	1.081	19.58	1.308	93.11	0.989	18.89	0.988	76.00	0.95	22.25	0.93	79.73	0.703	20.45	0.71
	New	68.00		21.00		92.56		18.78		69.45		21.30		62.00		18.08	
3	Old	109.83	1.045	42.42	0.983	102.11	0.977	37.11	1.041	104.58	0.994	45.64	0.915	120.09	0.977	39.27	1.13
	New	118.23		41.92		99.22		38.11		103.64		42.91		118.33		43.67	
4	Old	121.42	0.998	63.42	0.98	103.00	1.003	57.22	1.048	99.86	1.021	64.42	1.029	111.36	1.001	58.18	1.041
	New	120.85		62.54		103.44		60.33		102.45		65.91		111.58		60.67	
5	Old	182.25	0.955	125.67	0.931	158.44	0.924	103.67	0.95	170.64	1.031	122.50	1.037	175.27	0.95	106.55	1.06
	New	176.38		118.23		148.67		98.44		175.30		125.03		161.50		110.58	
6	Old	105.50	0.934	89.67	0.931	104.11	1.032	86.22	1.068	121.81	0.975	101.44	0.978	104.64	0.995	82.64	1.041
	New	96.23		81.69		109.22		91.67		117.12		98.94		104.08		85.58	
7	Old	105.00	1.081	93.33	1.011	110.11	1.004	92.78	1.042	122.31	1.014	105.03	1.034	113.09	1.077	93.91	1.122
	New	110.62		94.31		110.56		96.00		124.67		108.67		118.92		101.50	
8	Old	98.17	0.995	80.67	0.992	99.89	0.928	88.78	0.94	95.86	1.018	83.94	1.042	81.27	1.142	70.27	1.142
	New	97.46		79.92		95.11		84.78		98.76		88.45		89.08		79.67	
9	Old	62.92	1.023	57.92	1.025	73.67	1.027	68.56	1.031	70.06	1.004	65.44	1	65.64	1.078	57.27	1.221
	New	65.23		60.08		77.11		72.33		70.52		65.48		73.50		66.75	
10	Old	47.08	1.042	44.67	1.06	54.44	1.064	51.56	1.055	50.00	1.011	47.25	1.023	49.45	1.086	42.82	1.288
	New	49.92		48.15		56.56		53.33		50.73		48.73		55.42		51.58	
11	Old	27.50	1.048	26.25	1.066	36.33	1.255	34.78	1.188	27.69	0.995	26.44	0.993	28.09	1.422	26.18	1.375
	New	29.92		29.00		42.22		40.44		27.45		26.18		37.50		35.50	
12	Old	19.83	1.107	19.00	1.104	23.44	1.124	22.67	1.139	19.31	1.048	18.75	1.052	18.18	1.129	17.00	1.178
	New	22.77		21.15		25.78		25.11		20.36		19.79		22.08		21.25	
13	Old	12.17	0.845	12.00	0.844	12.44	1.103	11.89	1.086	9.97	0.995	9.64	1.013	10.82	1.074	10.36	1.078
	New	9.54		9.54		13.89		13.11		9.91		9.79		11.75		11.33	
14	Old	5.50	1.415	5.42	1.463	7.78	0.969	7.78	0.969	4.11	1.177	4.08	1.174	5.91	0.919	5.73	0.945
	New	7.15		7.08		7.67		7.67		5.06		5.03		5.42		5.42	
15	Old	3.92	1.066	3.83	1.059	3.33	1.38	3.22	1.405	1.97	1.067	1.97	1.056	2.09	1.278	2.09	1.278
	New	4.31		4.15		4.78		4.78		2.15		2.12		2.75		2.75	
16	Old	1.75	0.895	1.67	0.93	1.67	0.914	1.67	0.914	0.83	0.936	0.83	0.902	0.82	2.152	0.82	2.152
	New	1.54		1.54		1.56		1.56		0.76		0.73		1.83		1.83	
17	Old	1.17	1.289	1.17	1.289	0.89	1.168	0.89	1.168	0.69	0.96	0.69	0.96	1.00	0.75	1.00	0.75
	New	1.54		1.54		1.00		1.00		0.67		0.67		0.67		0.67	

(continued)

Table A2 Continued

		Panel B: For the works of Shakespeare											
		H				M				Y K			
		Tokens		Types		Tokens		Types		Tokens		Types	
Word Length	Work type	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio
1	Old	16.84	1.004	5.32	0.935	24.29	0.773	7.29	0.547	40.42	1.009	6.75	0.835
	New	17.04		5.16		14.29		5.00		43.33		5.58	
2	Old	84.72	1.206	24.56	1.389	80.00	1.011	25.00	0.92	83.08	1.068	25.33	0.916
	New	108.24		28.2		81.79		24.00		92.33		24.00	
3	Old	106.68	0.926	47.6	1.094	104.00	1.011	49.79	0.993	90.25	1.109	43.75	1.175
	New	100.64		50.56		105.64		49.50		99.17		48.17	
4	Old	110.64	1.134	71.08	1.045	124.64	0.993	71.43	0.94	111.92	1.044	67.83	1.07
	New	126.44		73.04		122.50		69.21		119.58		73.08	
5	Old	167.76	0.981	119.36	1.028	174.29	1.009	123.57	0.977	169.92	0.986	122.75	1.01
	New	164.32		122.48		176.29		122.36		167.67		123.67	
6	Old	118.36	1.009	87.64	1.02	111.36	1.003	89.43	0.98	127.17	0.999	94.33	0.973
	New	120.48		89.6		111.93		87.86		127.08		92.17	
7	Old	122.56	1.012	97.48	1.009	138.50	1.005	100.86	0.971	114.25	1.086	91.92	1.113
	New	126.2		98.68		139.57		97.29		124.08		99.58	
8	Old	95.28	0.922	81.84	0.906	82.64	0.976	76.43	0.985	85.08	0.87	76.33	0.914
	New	82.76		71.88		81.21		75.64		77.00		71.00	
9	Old	70	0.905	58.48	0.915	60.57	1.008	54.71	1.016	56.67	0.947	53.25	0.923
	New	57.28		52		61.21		55.57		51.83		47.83	
10	Old	43.32	0.909	41.48	0.887	40.43	1.018	39.14	1.01	56.83	0.962	41.00	0.96
	New	39.68		37.72		41.57		39.64		53.50		37.17	
11	Old	26.4	0.999	24	0.994	25.57	1.103	25.00	1.111	26.33	0.819	25.50	0.784
	New	26.32		23.72		28.07		27.50		20.50		19.33	
12	Old	19.4	0.927	17.64	0.895	16.64	1.013	16.50	1.013	17.25	0.852	16.92	0.838
	New	16.56		15.12		16.93		16.79		13.42		13.00	
13	Old	8.44	0.866	8.36	0.863	8.57	1.118	8.57	1.118	10.08	0.579	9.83	0.439
	New	6.76		6.64		9.57		9.57		5.50		5.25	
14	Old	4.24	0.741	4.16	0.745	4.07	1.026	4.07	1.026	5.33	0.801	5.33	0.801
	New	3		2.96		4.21		4.21		3.58		3.58	
15	Old	2.44	1.041	2.4	1.054	2.29	1.48	2.29	1.48	3.08	0.157	3.08	0.157
	New	2.6		2.6		3.21		3.21		0.50		0.50	
16	Old	1.32	0.82	1.32	0.82	1.57	0.824	1.57	0.824	1.50	0.266	1.50	0.266
	New	1.04		1.04		1.21		1.21		0.58		0.58	
17	Old	0.76	0.499	0.76	0.499	0.21	3.667	0.21	3.667	0.75	0.516	0.75	0.516
	New	0.44		0.44		0.50		0.50		0.33		0.33	

Table A3 Logistic regression results comparing token and type stem lengths between old and new translations

Panel A: For the works of Daudet, Dostoyevsky, and Montaigne

		B G				D				D M				U K			
		Tokens		Types		Tokens		Types		Tokens		Types		Tokens		Types	
Stem Length	Work type	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds Ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio
1	Old	42.50	0.947	2.75	1.123	29	0.924	3.67	0.839	28.89	0.924	5.11	0.695	53.09	0.869	4.82	0.1
	New	29.46		3.00		23.44		3.33		25.28		4.47		38.83		3.25	
2	Old	162.25	1.058	34.00	1.131	197.22	1.026	33.56	1.196	159.11	0.968	38.86	0.796	188.00	0.883	33.64	1.05
	New	171.08		35.69		200.00		34.56		149.41		36.75		163.67		34.00	
3	Old	313.42	1.045	94.25	1.035	276.44	1.025	88.44	1.057	277.78	1.006	107.58	0.955	282.64	1.135	84.91	1.676
	New	342.31		97.62		285.22		93.44		281.44		105.16		341.83		102.75	
4	Old	130.17	1.021	63.25	1.037	124.22	1.054	64.33	1.254	129.08	1.094	74.31	1.161	109.18	1.123	59.09	1.194
	New	139.08		66.54		138.56		72.44		147.69		84.56		130.08		65.25	
5	Old	205.25	1.013	108.17	0.981	233.67	1.008	110.89	1.04	211.89	1.045	126.14	1.031	218.55	0.986	100.36	1.089
	New	212.69		103.15		237.89		117.00		229.91		133.94		215.58		104.92	
6	Old	55.83	0.92	42.17	0.825	70.22	0.887	49.11	0.909	76.19	0.999	55.06	0.975	69.27	0.781	43.55	0.53
	New	49.31		33.31		55.89		41.89		75.97		53.13		56.75		34.75	
7	Old	35.83	0.783	27.58	0.756	43.11	0.96	29.78	0.931	49.50	0.981	35.89	0.919	42.64	0.829	26.82	0.834
	New	22.15		17.77		36.78		25.89		47.38		32.88		31.75		22.17	
8	Old	28.42	0.887	15.25	0.494	11.56	0.958	10.44	0.973	30.89	0.95	21.72	0.792	14.64	0.744	11.09	0.568
	New	19.69		8.00		10.67		10.00		23.75		14.91		9.67		7.33	
9	Old	6.92	0.765	6.42	0.633	6.22	0.525	5.56	0.455	12.22	0.657	10.53	0.571	8.00	0.749	6.09	0.52
	New	3.62		2.85		3.89		3.33		7.41		5.91		3.33		2.00	
10	Old	8.00	0.61	6.75	0.47	3.56	1	3.00	0.963	10.11	0.585	8.61	0.461	7.82	0.741	4.18	0.792
	New	4.31		3.69		3.56		2.89		4.94		4.38		4.17		3.17	
11	Old	3.50	0.294	3.25	0.284	1.67	1.191	1.67	1.191	5.17	0.537	4.61	0.381	2.00	0.549	1.91	0.556
	New	1.00		1.00		1.89		1.89		2.94		2.31		1.25		1.17	
12	Old	2.00	0.851	2.00	0.813	1.33	0.239	1.33	0.239	4.28	0.464	4.14	0.358	1.64	0.814	1.45	0.835
	New	1.62		1.54		0.44		0.44		1.81		1.63		1.25		1.17	

(continued)

Table A3 Continued

		Panel B:For the works of Shakespeare											
		H				M				Y K			
		Tokens		Types		Tokens		Types		Tokens		Types	
Stem Length	Work type	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio	Avg. Occ.	Odds ratio
1	Old	27.8	0.993	4.72	0.754	34.50	0.826	5.86	0.399	55	1.005	6.83	0.598
	New	27.28		4.04		26.14		3.64		56.41667		4.67	
2	Old	179.56	1.091	38.8	1.064	176.00	1.002	40.14	0.933	180.5833	1.036	37.42	0.941
	New	197		39.64		177.07		39.07		191.5833		36.50	
3	Old	272.64	1.04	96	1.066	291.50	1.067	102.93	1.05	252.3333	1.153	91.50	1.083
	New	285.8		101.12		311.07		105.07		296.6667		96.33	
4	Old	112.8	1.246	66.4	1.124	132.64	1.011	70.00	0.937	113	1.035	65.25	1.045
	New	148.16		72.72		135.57		68.14		122.5833		67.42	
5	Old	187.56	0.978	107.64	1.028	199.43	1.005	112.93	0.934	206.0833	0.889	115.75	0.889
	New	181.8		111.68		201.50		108.00		176		100.17	
6	Old	88.96	0.943	48.96	0.802	72.79	0.979	46.29	0.849	91.83333	0.92	50.17	0.751
	New	71.4		39.44		68.64		40.43		80.33333		38.58	
7	Old	53.2	0.978	32.68	0.809	65.43	0.94	30.93	0.705	41.33333	0.937	27.92	0.751
	New	49.44		25.6		59.93		22.43		34.58333		21.58	
8	Old	31.4	0.851	21.2	0.339	12.64	0.807	8.43	0.679	15.5	0.774	13.25	0.493
	New	16.56		9.76		9.50		6.43		11.83333		8.50	
9	Old	16.52	0.835	12.72	0.366	7.21	0.872	4.71	0.79	10.33333	0.754	8.42	0.557
	New	9.16		5.32		5.29		3.64		4.916667		3.75	
10	Old	9.24	0.583	8.52	0.52	2.64	0.71	2.50	0.672	22.16667	0.95	7.42	0.572
	New	3.68		3.08		1.86		1.71		18.83333		3.92	
11	Old	7.8	0.892	5.72	0.492	1.43	0.838	1.43	0.838	5	0.809	4.58	0.604
	New	5.08		2.92		1.14		1.14		3.333333		2.25	
12	Old	6.92	0.749	5.32	0.3	1.50	0.377	1.43	0.374	2.25	0.72	2.25	0.553
	New	3		1.72		0.71		0.71		1.25		1.00	