transLogos 2021 Vol 4 Issue 1
Bozan, Nalan, pp. 57–80
https://dx.doi.org/10.29228/transLogos.31

transLogos
A Translation Studies Journal

A Text Mining Approach to 'Operational Norms'

Nalan BOZAN[*]

The notion of 'translation norms' was introduced by Gideon Toury in the late 1970s within Descriptive Translation Studies (DTS). Toury's endeavor was to describe and explain translations to make generalizations regarding translational behavior. Due to the exhaustive nature of such attempt, analysis of everything is deemed 'untenable' by Toury, and even sampling is limited to the capacity of human effort. This paper[1] argues that exhaustive analysis on 'operational norms' can be facilitated through text mining. Text mining is a term that can be described as automated analysis of text data aided by a software to extract and discover new insights. As part of this study, a custom-made software tool is designed, and the corpus of a master's thesis (Duymaz 2020) is used as a reference point to analyze and compare the findings with a view to testing the effectiveness of the software versus human effort. The tool is designed by a software expert, and then the test data is subjected to a preparation stage to achieve optimal results by the software. After processing the test data, operational norms are extracted to be interpreted. The analysis has provided results that cannot be obtained by sole human effort, enabling researchers to have further insights about translations. This research thus has a potential for making a notable contribution to DTS. Further studies will improve the tool verifying its viability for various text types to ensure such contribution.

Keywords: translation norms; operational norms; text mining; Descriptive Translation Studies (DTS); exhaustive analysis

1. Introduction

Gideon Toury underlines the importance of descriptive branch in a discipline and mentions "the descriptive-explanatory goal of supplying exhaustive accounts of whatever has been presented/regarded as translational within a target culture" (2012, 20). Yet he deems analysis of everything "untenable" and mentions the difficulties in supplying sampling rules owing to human limitations (71). Even though no strict statistical methods are pointed out in his work, Toury does not rule out the advances in computer sciences which may allow much more and bigger samples (91–92). Previous studies show the significant contribution of using computerized corpora and text mining. However, to this author's knowledge, no special focus

[*] Master's student at Istanbul 29 Mayıs University.
E-mail: bozann18@29mayis.edu.tr; ORCID ID: https://orcid.org/0000-0002-1286-7029.

[1] This paper is derived from the ongoing research for the author's master's thesis on the same subject at Istanbul 29 Mayıs University.

(Received 11 January 2021; accepted 18 June 2021)

has been placed on 'translation norms.' This research therefore aims to facilitate exhaustive analysis on 'operational norms' through text mining.

In line with the research's aim, a software tool is designed. This paper seeks to determine whether the tool—or text mining—facilitates the exhaustive analysis, and if so, how it achieves it. Also, capabilities and limits of the tool will be tested. Whether the tool is suitable for all types of texts is another research question that needs to be answered. As the software is designed for analysis only on 'operational norms,' the reason behind this choice will be explained in the paper. Finally, this study aims to address the ultimate question on how this study contributes to descriptive translation analysis.

2. Background and Conceptual Framework

The ground for this research is based on the key concept and the theoretical framework introduced by Toury (1978, 1995, 2012). Also, the research proposes text mining as a tool to be utilized in translation analysis. The concepts are explicated below.

2.1 Operational Norms

Toury (2012) states that the main goal of Translation Studies (TS) is to describe, explain, and predict translation phenomena to become an autonomous discipline, and he provides a research methodology under the umbrella of 'Descriptive Translation Studies' (DTS).

A key concept in DTS is "norm," which was introduced in 1978 by Toury with regard to the regularities of translational behavior. He points out the cultural significance of translation activities and the social role that translator plays. This socio-cultural dimension of translation involves "norms." Drawing on the definition used in sociology, Toury defines norms as social values shared by a community in regard to appropriate behavior and refers to translation as a norm-governed activity. It should be noted that the notion of norm here does not indicate a prescriptive category but a descriptive one. That is, the focus is not to evaluate translations but to investigate the regularities of translation behavior.

Toury (1995, 56–61) suggests three types of norms: "initial norm," "preliminary norms," and "operational norms." The initial norm refers to the basic choice that a translator makes to subject him-/herself either to the original text—and its norms—or to the norms that are active in the target culture. Adherence to source norms determines a translation's "adequacy" whereas subscription to the norms of target culture determines its "acceptability." The second type of norms is preliminary norms, which deal with two interconnected considerations: "translation policy," which refers to the choice of text types, individual texts, human agents and groups, etc., and the "directness of translation," which concerns the threshold of tolerance for indirect translation. The third type is operational norms. They involve the decisions that the translator makes during the act of translation, and they are divided into two further types: "matricial norms" and "textual-linguistic norms." Matricial norms are related to the completeness of translation, distribution of text, and segmentation of target-text material. On the other hand, textual-linguistic norms concern the selection of material to build the target text or replace material in original text with.

Toury (2012) also mentions difficulties in the establishment of sampling rules for translational behavior or the results of such behavior. He argues that an absolute consideration of everything is "untenable" (71). Nevertheless, he does not rule out the possibility of advances in computing world which may enable more and much bigger samples and remarks that "much energy should still be directed toward the crystallization of systematic research methods, including quantitative ones" (92). From the point of the current research, such possibility stimulates the search for new research methods or tools.

Another theoretical motivation for this research can be found in Mona Baker's study (1993). It discusses the impact of availability of corpus on the study of translation with reference to some of the applications of corpus techniques in the applied branch of TS. She argues that "the potential for using large computerised corpora in translation studies" needs to be explored for the development of descriptive branch of TS. Although the corpus-driven aspect of her study is not directly relevant to this research, Baker's approach is influential as it calls for new techniques and a set of tools "to develop a descriptive branch of the discipline with well-defined objectives and an explicit program" (248).

2.2 Text Mining

The other key concept that this research is built around is 'text mining.' This concept can be described as "function of software or hardware components in a computer system which analyze or synthesize spoken or written language" (Jackson and Moulinier 2002, 2–3). As Nadir Zanini and Vikas Dhawan (2015) explain, main goal in text mining is to "turn text into data so that it is suitable for analysis” because the process involves discovery of hidden information in unstructured texts.

Chunyu Kit and Jian-Yun Nie (2015) state that “text mining is considered a variation or extension of *data mining* (DM), which is also known as knowledge discovery in database (KDD)” and note a difference between text mining and data mining: “Instead of working on databases, [text mining] works on unstructured texts” (510). Unstructured data requires a preparation stage so that algorithms can work on them. Then text mining tasks can be performed which are including but not limited to ‘information extraction,’ ‘text categorization,’ ‘text clustering,’ ‘summarization,’ and ‘sentiment analysis.’

We can see various applications of text mining in different areas like social media platforms (e.g., understanding target audience’s reactions), marketing (e.g., analyzing customer behavior and identifying trends), customer care services (e.g., automated analysis of customer feedback), insurance (e.g., fraud detection), academia (e.g., plagiarism detection), and so on.

A closely related concept is ‘natural language processing’ or NLP. Basically, it helps computers to understand and handle human languages. Olivia Kwong Oi Yee (2015, 563) points out “two major goals” of NLP: (i) “to enable human-computer interaction with human languages as the medium,” (ii) “to build language application systems which require considerable human language abilities and linguistic knowledge.” Text mining systems use several NLP techniques like ‘tokenization,’ ‘lemmatization,’ ‘parsing,’ ‘stemming,’ and ‘stop removal’ in order to build inputs for machine learning models.

3. Literature Review

There have been a number of studies conducted under corpus-based TS which employ similar processing methods. Yet, to this author’s knowledge, no study has been explicitly driven by the ‘operational norms’ as defined within Toury’s study (1978, 1995, 2012). Also, they are aimed to fulfil various different purposes.

Baker (2000) offers a corpus for stylistic analysis of literary translation outlining a methodological framework for investigating a literary translator’s style. The corpus which is named as “Translational English Corpus” is comprised of English texts translated from a variety of source languages, both European and non-European. Also, a software tool was developed to process the corpus semi-automatically. In order to investigate the style of individual literary translators, she identifies some patterns related to "type/token ratio; average sentence length; variation across texts; frequency and patterning of SAY (the most frequent reporting verb in English)" (249). Revisiting theoretical discussions around the translator's voice and 'visibility' (see Venuti 1995) on whether translation is a creative or a reproductive activity, Baker embarks on a quest for translator's "individual imprints" on the produced text by using a tool which inherently uses text mining techniques.

A further example of corpus-assisted translator style studies was carried out by Defeng Li (2017) incorporating meta- and paratexts into the analysis. He first explains the discussion on style in translation and then reviews the existing corpus-assisted translator style studies and discusses their methodological issues. He argues for "an integration of quantitative and qualitative analysis" and contends that "the qualitative analysis of examining the translator's style in a larger socio-cultural context remains the ultimate goal" (103), illustrating his argument with a case study in which he compares English translations of the classical Chinese novel *Hongloumeng*.

Tom Cheesman et al. (2017) conduct an experimental project developing "a web-based system which enables users to create parallel, segment-aligned multi-version corpora" and revealing variation in multiple translations through visual interfaces. This project provides an important opportunity to understand how digital tools "could be developed to explore patterns in variation among (re)translations" (743).

Johanna Monti et al. (2011) aim to explore a new approach based on a text mining tool in order to help "translators look for different types of information (glossaries, corpora, Wikipedia, and so on) related to the specific translation work that they have to perform which can then be used to update the lexical base needed for the translation workflow."

There are two studies conducted in computer sciences which ultimately aim to test the efficiency or limit of text mining and/or NLP while using translation as an object of research. The first one is "Determining Translation Invariant Characteristics of James Joyce's *Dubliners*" by Jon M. Patton and Fazlı Can (2012). The researchers perform a comparative stylometric analysis using five style markers: (i) sentence length in terms of the number of words, (ii) the most frequent words, (iii) word length in text, (iv) word length in vocabulary, and (v) vocabulary richness. In other words, they use statistical methods to analyze and compare writing styles of the author and the translator. Patton and Can explain their aim as "to identify style-related features of the original work which are retained in translation" (210).

The second study is Sevilay Çalışkan's (2020) master's thesis entitled "Text Mining Analysis of Translation, Social Communication and Literary Writing for Turkish." She "analyze[s] different types of Turkish text from different points of views, having an overall review on text mining in Turkish at the end" (1). Her research consists of four parts, but only the first part is relevant to this paper as it involves translations. The other parts include either monolingual analysis of different text types or multilingual analysis of different texts in the same domain. In the first part of Çalışkan's research, "loyalty of translations" for the Turkish novel *Benim Adım Kırmızı* (*My Name is Red*) into English, French, and Spanish is evaluated. The evaluation is based on the hypothesis that "the styles of characters will be preserved to some point in the translations if the style has not [been] changed consciously by translator" (3).

4. Methodology

The research is carried out in four stages:

1. Designing the tool
   a. Defining categories according to 'operational norms'
   b. Developing the software
2. Preparing test data
3. Processing test data
4. Extracting and analyzing operational norms

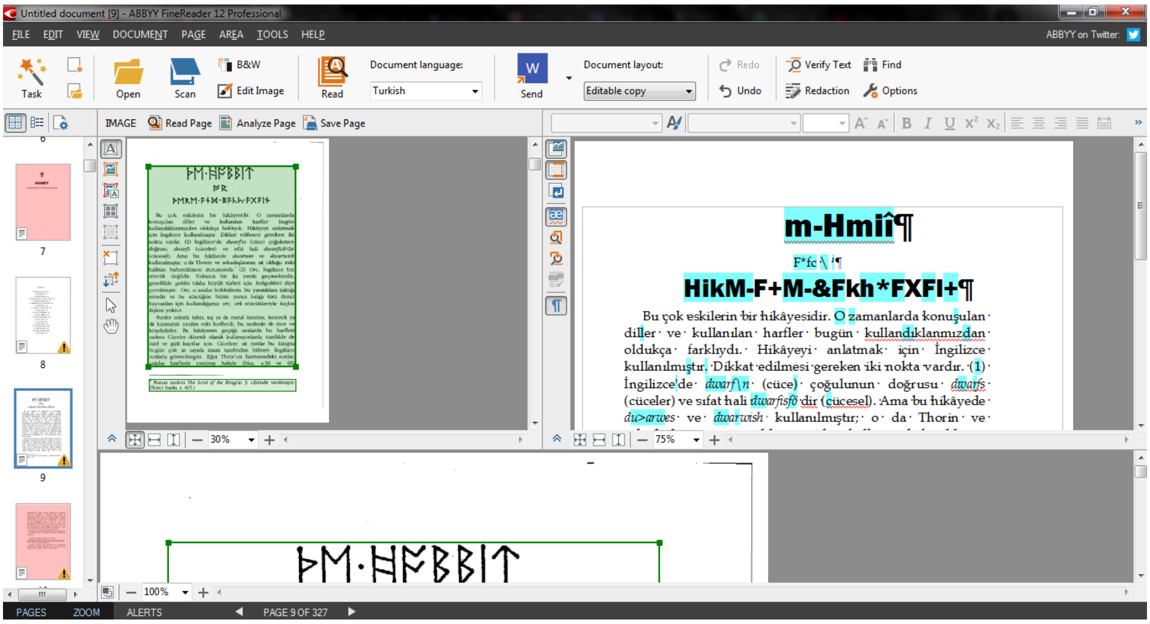
First stage involves the design and development of the tool entailing interdisciplinary cooperation. Software of the tool is developed by a software expert[2] outside the field of TS. For this reason, the demanded features or functions are reported to the expert based on the aspects/factors which are investigated within the framework of operational norms.

At the second stage, a set of test data is converted to computer-readable digital format. As the process is demonstrated in figure 1, this stage includes scanning of printed texts and their optical character recognition (OCR). For OCR, ABBYY FineReader is used, and the resulted text is edited whenever it is necessary (see figure 2). Then the texts are exported into HTML format. In this way, hyper-text markup language makes it possible to tag the paragraphs, images, footnotes, superscripts, and so on. This can be considered as defining the components

[2] Kerim Bozan (E-mail: kerim@bozan.biz).

of each text and hence introducing them to the software. The key point of this stage is to provide the most flawless version of a document. In other words, the better optically recognized a scanned document is, the more accurate tags are obtained.

Figure 1. Optical character recognition (screenshot from ABBY FineReader showing the optical character recognition process of target text 1)



Document language: Turkish

Document layout: Editable copy

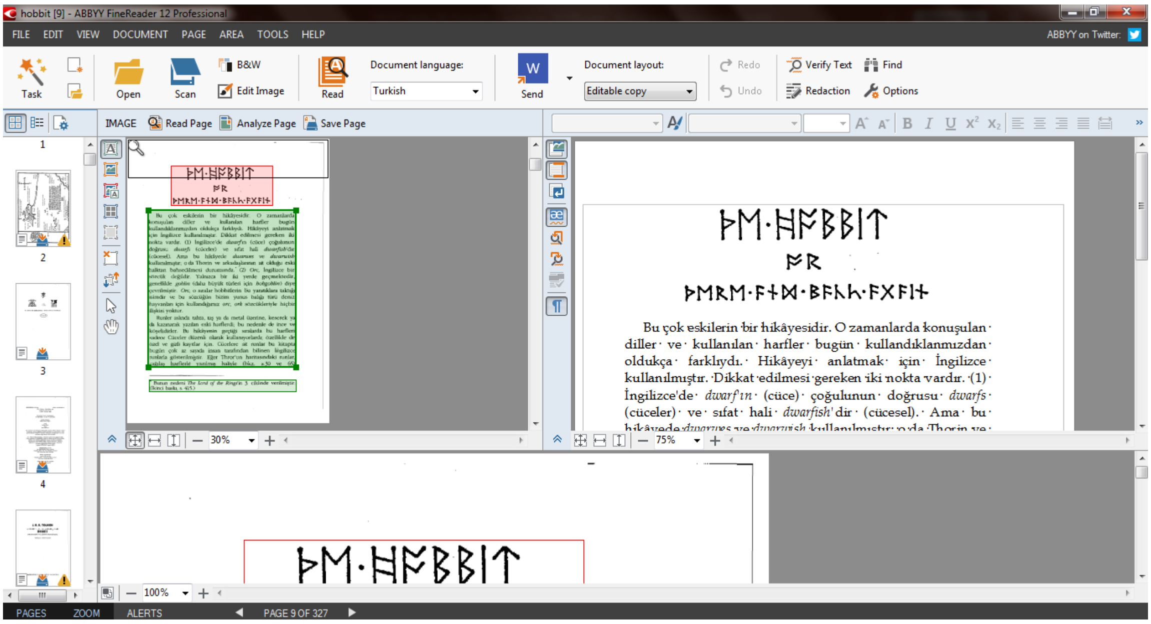
m-Hmiî¶

F*fc \ '¶

HikM-F+M-&Fkh*FXFI+¶

Bu çok eskilerin bir hikâyesidir. O zamanlarda konuşulan diller ve kullanılan harfler bugün kullandıklanmızdan oldukça farklıydı. Hikâyeyi anlatmak için İngilizce kullanılmıştır. Dikkat edilmesi gereken iki nokta vardır. (1) İngilizce'de *dwarf\n* (cüce) çoğulunun doğrusu *dwarfs* (cüceler) ve sıfat hali *dwarfisfö* dir (cücesel). Ama bu hikâyede *du>arwes* ve *dwarwish* kullanılmıştır; o da Thorin ve

Figure 2. Editing process (screenshot from ABBY FineReader showing the edited version of the optically recognized target text 1)



Document language: Turkish

Document layout: Editable copy

Bu çok eskilerin bir hikâyesidir. O zamanlarda konuşulan diller ve kullanılan harfler bugün kullandıklanmızdan oldukça farklıydı. Hikâyeyi anlatmak için İngilizce kullanılmıştır. Dikkat edilmesi gereken iki nokta vardır. (1) İngilizce'de *dwarf'ın* (cüce) çoğulunun doğrusu *dwarfs* (cüceler) ve sıfat hali *dwarfish'* dir (cücesel). Ama bu

Subsequent stage deals with the processing of test data using the tool designed in the first stage. The software processes the texts and presents the source text and the translated text(s) through a user-friendly interface.

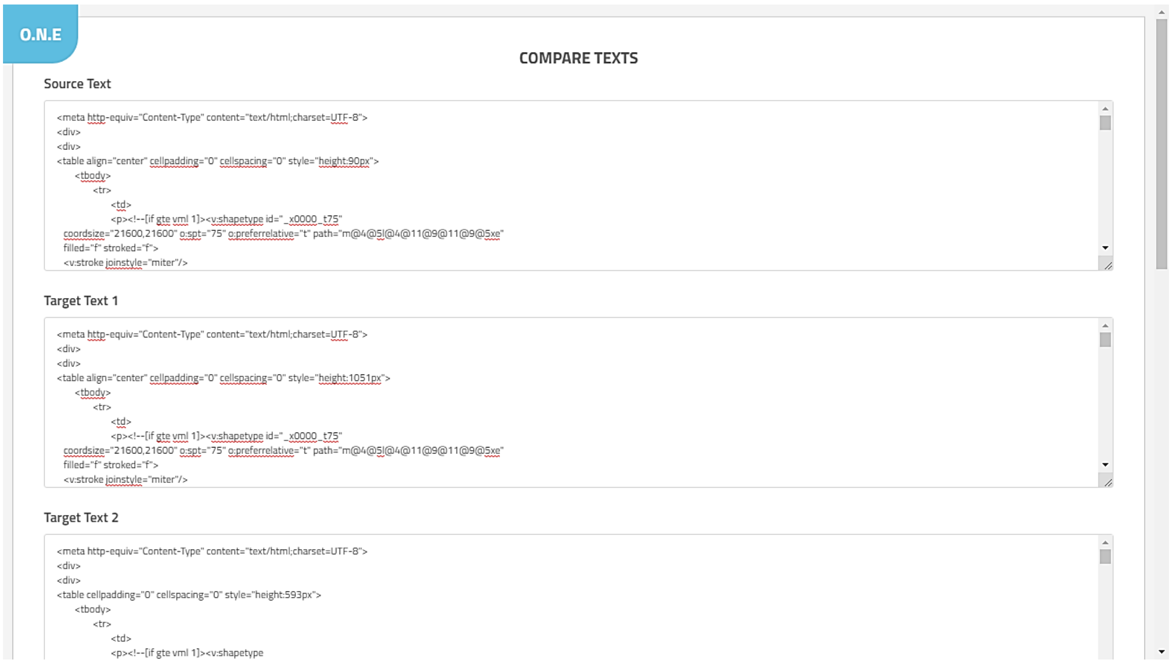
Last stage provides the results of processing under the categories specified within the framework of operational norms. Visualization of the results is maintained through graphs and/or tables, which helps the researcher to analyze operational norms.

4.1 Tool: O.N.E

O.N.E (Operational Norm Extractor) is a custom-made tool involving text mining techniques. It is a browser-based software tool. Thus, one can upload as many texts as they wish to analyze. However, it should be noted that it depends on the capacity of their servers. In other words, the more texts are uploaded, the more time it will take the tool to process them.

The texts are imported into the tool in HTML format, and the user can view the texts in the same frame one below the other (see figure 3). Analyzed texts are followed by the checkboxes of the categories which the analysis will be based on (see figure 4).

Figure 3. User interface of O.N.E – texts section (screenshot of the texts section of O.N.E)



Source Text

```
<meta http-equiv="Content-Type" content="text/html;charset=UTF-8">
<div>
<div>
<table align="center" cellpadding="0" cellspacing="0" style="height:90px">
    <tbody>
        <tr>
            <td>
            <p><!--[if gte vml 1]><v:shapetype id="_x0000_t75"
  coordsize="21600,21600" o:spt="75" o:preferrelative="t" path="m@4@5l@4@11@9@11@9@5xe"
  filled="f" stroked="f">
  <v:stroke joinstyle="miter"/>
```

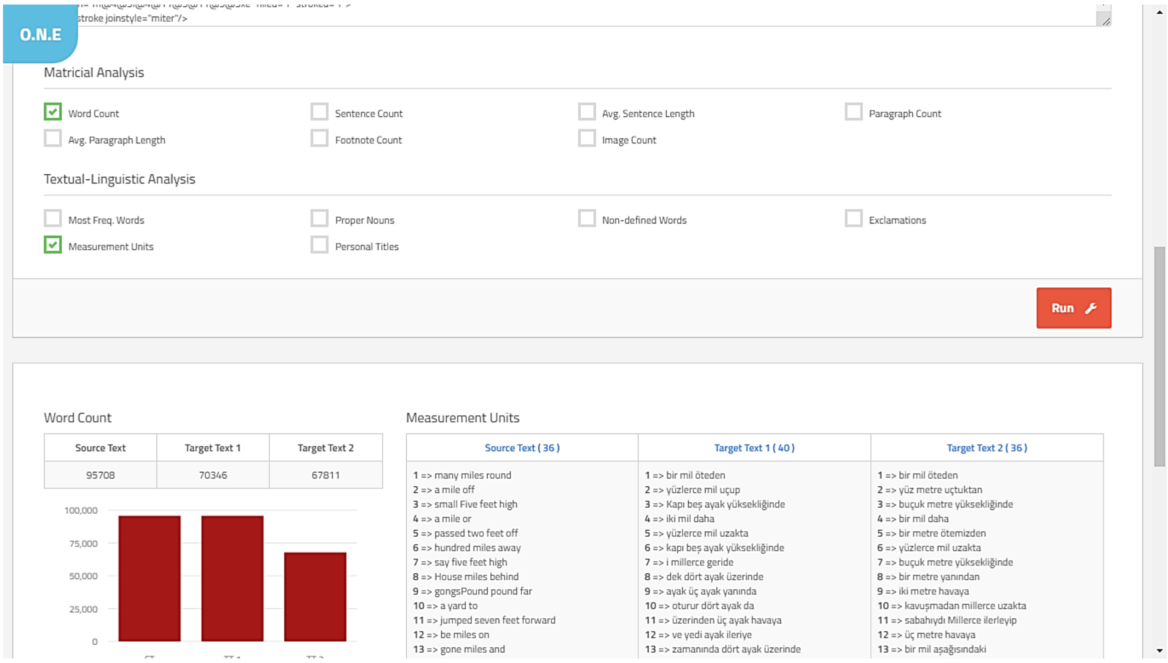
Target Text 1

```
<meta http-equiv="Content-Type" content="text/html;charset=UTF-8">
<div>
<div>
<table align="center" cellpadding="0" cellspacing="0" style="height:1051px">
    <tbody>
        <tr>
            <td>
            <p><!--[if gte vml 1]><v:shapetype id="_x0000_t75"
  coordsize="21600,21600" o:spt="75" o:preferrelative="t" path="m@4@5l@4@11@9@11@9@5xe"
  filled="f" stroked="f">
  <v:stroke joinstyle="miter"/>
```

Target Text 2

```
<meta http-equiv="Content-Type" content="text/html;charset=UTF-8">
<div>
<div>
<table cellpadding="0" cellspacing="0" style="height:593px">
    <tbody>
        <tr>
            <td>
            <p><!--[if gte vml 1]><v:shapetype
```

Figure 4. User interface of O.N.E – analysis section (screenshot of the analysis section of O.N.E)



```
stroke joinstyle="miter"/>
```

Matricial Analysis

☑ Word Count
□ Sentence Count
□ Avg. Sentence Length
□ Paragraph Count
□ Avg. Paragraph Length
□ Footnote Count
□ Image Count

Textual-Linguistic Analysis

□ Most Freq. Words
□ Proper Nouns
□ Non-defined Words
□ Exclamations
☑ Measurement Units
□ Personal Titles

Word Count

Source Text	Target Text 1	Target Text 2
95708	70346	67811

Measurement Units

Source Text (36)	Target Text 1 (40)	Target Text 2 (36)
1 => many miles round 2 => a mile off 3 => small Five feet high 4 => a mile or 5 => passed two feet off 6 => hundred miles away 7 => say five feet high 8 => House miles behind 9 => gongsPound pound far 10 => a yard to 11 => jumped seven feet forward 12 => be miles on 13 => gone miles and	1 => bir mil öteden 2 => yüzlerce mil uçup 3 => Kapı beş ayak yüksekliğinde 4 => iki mil daha 5 => yüzlerce mil uzakta 6 => kapı beş ayak yüksekliğinde 7 => i millerce geride 8 => dek dört ayak üzerinde 9 => ayak üç ayak yanında 10 => oturur dört ayak da 11 => üzerinden üç ayak havaya 12 => ve yedi ayak ileriye 13 => zamanında dört ayak üzerinde	1 => bir mil öteden 2 => yüz metre uçtuktan 3 => buçuk metre yüksekliğinde 4 => bir mil daha 5 => bir metre ötemizden 6 => yüzlerce mil uzakta 7 => buçuk metre yüksekliğinde 8 => bir metre yanından 9 => iki metre havaya 10 => kavuşmadan millerce uzakta 11 => sabahıydı Millerce ilerleyip 12 => üç metre havaya 13 => bir mil aşağısındaki

4.2 Test Data

As mentioned earlier, the aim is to design a tool that helps researchers to analyze the operational norms by providing a larger amount of data. Nevertheless, it would not be possible to test all types of texts and present the results within a limited time. Therefore, this research would be an initial step towards further efforts to develop the tool. For the reasons explained here, the corpus of this research is limited to two different translations of the fantasy novel *The Hobbit* which was written by John Ronald Reuel Tolkien in 1937. The reason behind this specific choice is related to the possibility of comparing the results produced by the tool to the ones that have already been produced by human effort.

In her master’s thesis entitled “John Ronald Reuel Tolkien’in *The Hobbit* Adlı Eserinin Türkçe Çevirilerinin Yeniden Çeviri Varsayımı ve Çocuk Edebiyatı Kapsamında İncelenmesi” (An analysis of the Turkish translations of John Ronald Reuel Tolkien’s *The Hobbit* within the framework of retranslation hypothesis and children’s literature), Zeynep Duymaz (2020) analyzes *The Hobbit* within the framework of DTS creating a common ground or a reference point for comparison. It should be noted that such comparison is not intended to evaluate the quality of findings but rather to explore the limits or capacity of the proposed tool. In table 1, the source text and the target texts that Duymaz includes in her analysis are provided.

Table 1. Corpus of Duymaz's (2020) research

	Title	**Year**	**Publisher**	**Translator**
Source Text	*The Hobbit*	1990	HarperCollins	—
Target Text	*Hobbit: Oradaydık ve Şimdi Buradayız*	1996	Altıkırkbeş & Mitos	Emel İzmirli
	Hobbit: Oradaydık ve Şimdi Buradayız	1997	Altıkırkbeş	Esra Uzun
	Hobbit	2007	İthaki	Gamze Sarı
	Hobbit	2018	İthaki	Gamze Sarı

In the present research, e-pub edition of the source text, *The Hobbit* (Tolkien 2009), which will be referred to as 'Source Text' hereafter, was processed by the tool, along with scanned version of Emel İzmirli's translation, *Hobbit: Oradaydık ve Şimdi Buradayız* (Tolkien 1996)—'Target Text 1'—and e-book version of Gamze Sarı's translation (Tolkien 2015)—'Target Text 2.'

5. Findings

As stated earlier, the analysis involves another research conducted by Duymaz (2020) as a reference point to compare the results obtained by the tool with the ones obtained by human effort. Such comparison requires a closer look into the relevant categories analyzed in Duymaz's research. The categories in her analysis are given in table 2 vis-à-vis the categories in this analysis.

Table 2. Analyzed categories[3]

Duymaz's (2020) Research	**Current Research**
Matricial Norms	
—	Word count
	Sentence count
	Average sentence length
	Paragraph count
	Average paragraph length
	Footnote count
	Image count
Textual-Linguistic Norms	
Translation of proper nouns	Proper nouns
Translation of new words	Non-defined words
Translation of pun	—
Translation of idioms	—
Translation of colloquialisms and dialects	—
Translation of exclamations	Exclamations
Translation of food and beverage names	—
Translation of games	—
Translation of measurement units	Measurement units
—	Personal titles
—	Most frequent words

As shown in table 2, there are some differences in the categories, i.e., some categories are not included in this research, or they are presented differently. For instance, matricial norms are not presented in a structured way in Duymaz's research but rather described in paragraphs. In this research, on the other hand, quantitative and statistical data are focused based on various levels such as character, word, sentence, and paragraph. This difference is justifiable as it lies in the different nature and aims of these two studies. The other difference can be seen in the

[3] No subcategories of matricial norms are specified in Duymaz's paper. However, details such as title of the texts, content of the covers, number of pages, contents (forewords, images, etc.), number of chapters and how they are presented, length of the texts are explained in paragraphs and then compared.

categories under ‘textual-linguistic norms.’ Since Duymaz’s research involves the analysis of expressive language in the text and the translation of cultural elements, not all the categories could be included in this research. Yet, the author of this study believes it will be possible in the future when further studies are conducted with the help of NLP.

The results obtained from the analysis made by O.N.E will be presented under the above-given categories. Screenshots taken from the tool will be added wherever it is necessary. It should be noted that front cover texts (title, author, publisher, translator), table of contents and copyright page, ‘about the author’ or ‘about the book’ page(s), and epigraphs or the parts that are essentially different due to different publishers in general are manually removed from the analysis as they will affect the number of sentences and paragraphs in a way that prevents an accurate analysis.

5.1 Matricial Norms

The set of analyses that will be presented here examined the ‘matricial norms’ in the test data.

5.1.1 Word Count. Referring back to Duymaz’s research, length of the texts is measured based on the number of pages which changes depending on the style of the document, i.e., margins, font size, and paratexts. Duymaz (2020) herself is also aware of it as she explains that the longest text is the one that was published using the largest font (38). The present study, on the other hand, used word count as a criterion to measure the length as it can be seen in table 3.

Table 3. Word count

Source Text	**Target Text 1**	**Target Text 2**
95708	70346	67811

Under this category, a researcher can see the difference in the length between texts and try to justify the reason behind it. For example, the significant difference between the word counts of the source text and the target texts may be explained by the structural difference in languages, or the difference between the word counts of target texts may signal omission of some parts. To interpret the data better and reach a more accurate conclusion, the number of sentences and average paragraph length can be analyzed.

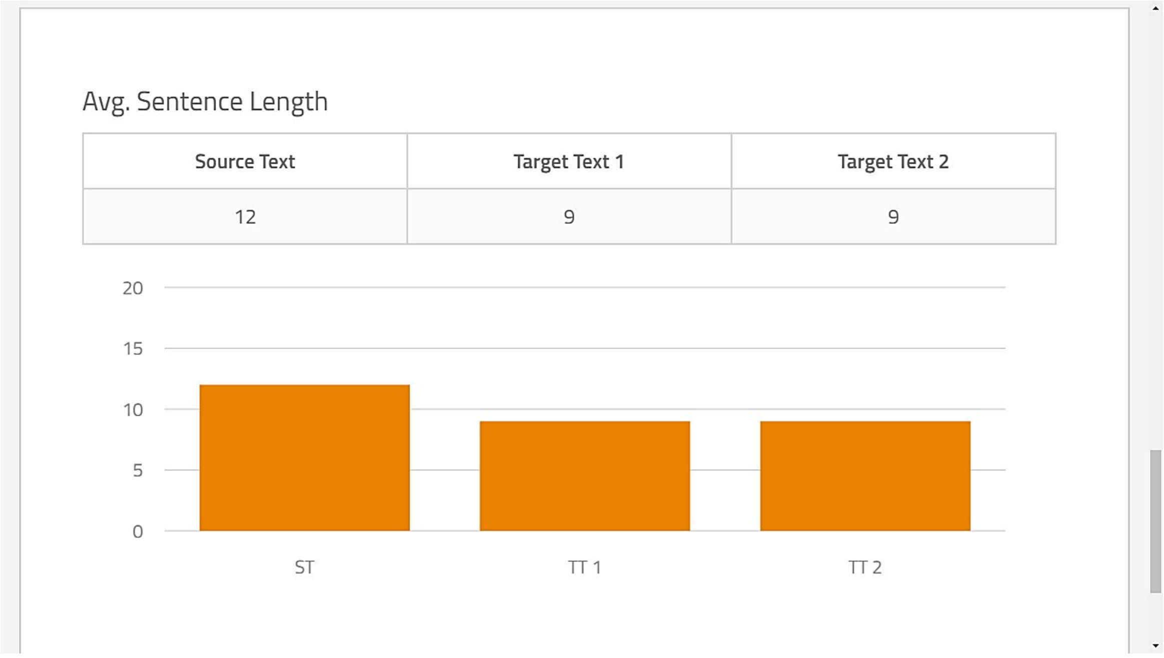
5.1.2 Sentence Count. It is apparent from table 4 that the gap between the word counts of Source Text and Target Text 1 is closed when the sentence count is analyzed. In other words, it can be said that the difference between word counts may be due to the difference between English and Turkish languages. However, there is still difference between Source Text and Target Text 2, and Target Text 1 and Target Text 2. It may be caused by omission of some paragraphs or a chapter. To support this possible cause, the researcher can compare the average sentence and paragraph lengths.

Table 4. Sentence count

Source Text	**Target Text 1**	**Target Text 2**
7954	7941	7730

5.1.3 Average Sentence Length and Average Paragraph Length. Average sentence length is calculated dividing total number of words by total number of sentences, while average paragraph length is calculated dividing total number of sentences by total number of paragraphs. The results are presented in figure 5 and figure 6.

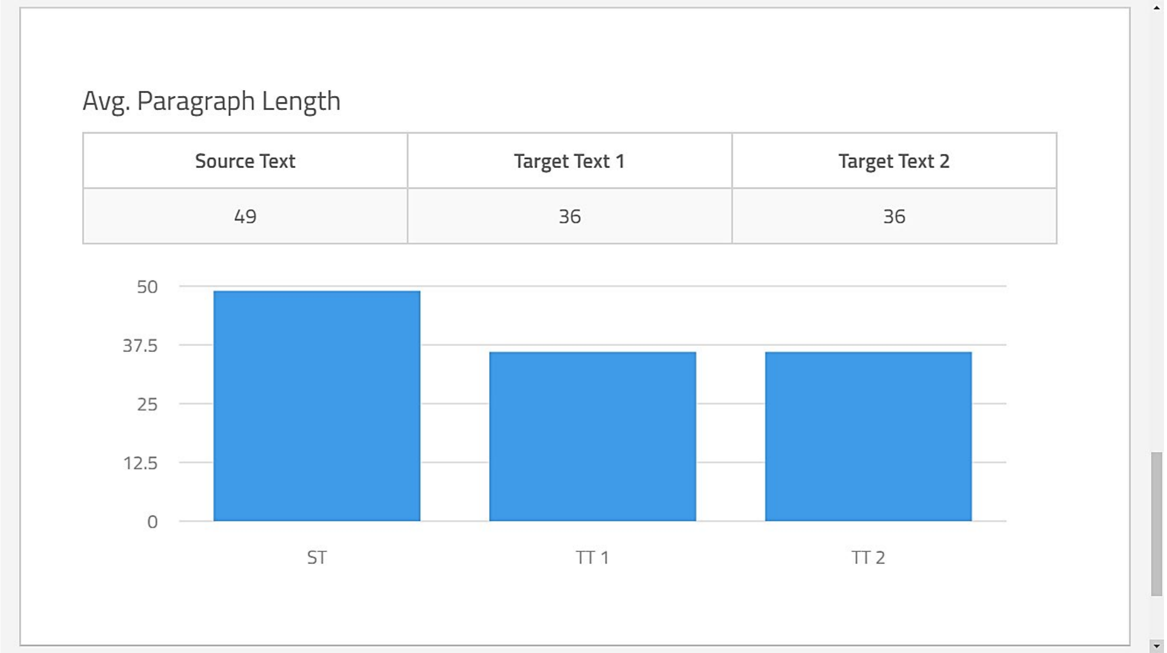
Figure 5. Average sentence length (screenshot from O.N.E showing the average sentence length; ST: Source Text, TT 1: Target Text 1, TT 2: Target Text 2)



Avg. Sentence Length

Source Text	Target Text 1	Target Text 2
12	9	9

Figure 6. Average paragraph length (screenshot from O.N.E showing the average paragraph length; ST: Source Text, TT 1: Target Text 1, TT 2: Target Text 2)



Avg. Paragraph Length

Source Text	Target Text 1	Target Text 2
49	36	36

The figures support the possible cause of the difference in word and sentence counts mentioned in the previous sections. Although the difference between Target Text 1 and Target Text 2 is smaller when compared to the difference between Source Text and Target Texts, it is high enough to further research the underlying cause, i.e., whether there is omission in Target Text 2 or addition to Target Text 1. As a matter of fact, a manual check confirms that ‘Author’s Note’ part is omitted in Target Text 2.

As for the major difference between Source Text and Target Texts, it may indicate major omissions in target texts or can be justified with structural difference in languages. Manual analysis confirms the latter possibility.

5.1.4 Footnote Count. The texts that are tagged as footnote are counted, and the results are provided in table 5. It can be seen that there is one more footnote in Target Text 2. This can be a signal for the researcher to further analyze footnotes. The extra footnote in Target Text 2 can be checked.

Table 5. Footnote count

Source Text	**Target Text 1**	**Target Text 2**
2	2	3

5.1.5 Image Count. Both Source Text and Target Texts contain images as the novel *The Hobbit* contains several maps and illustrations, yet there is difference in image counts of the text as it can be seen in table 6. These results can lead the researcher to ask why there are less images in Target Text 2.

Table 6. Image count

Source Text	**Target Text 1**	**Target Text 2**
19	19	5

5.2 Textual-Linguistic Norms

The categories under 'textual-linguistic norms' deal with the selection of material in the texts. Therefore, lexical databases for both English[4] and Turkish[5] are used in the tool.

5.2.1 Proper Nouns. The first category under 'textual-linguistic norms' compares the proper nouns in the texts, hence their translation. The number of words that are identified as proper noun by the tool is given in table 7.

Table 7. Proper nouns

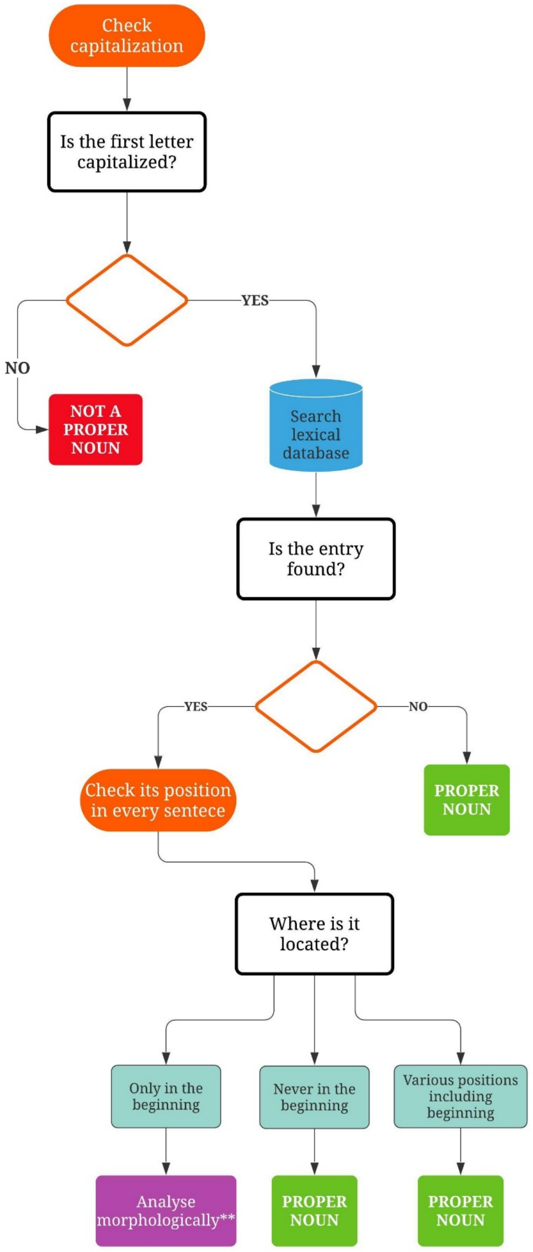
Source Text	**Target Text 1**	**Target Text 2**
278	487	491

Proper nouns are identified according to the process shown in figure 7. Capitalization in the first letter is an indicator of a proper noun, yet it is not sufficient since all the first letters in the beginning of a sentence are capitalized, including common words, adjectives, or other types of words. After the tool first examines capitalization, it refers to the lexical database to eliminate common words and other words except proper nouns. This process substantially

[4] Princeton University, "WordNet 3.1 Database," *WordNet*, 2010.
[5] "Frequency Effects in the Processing of Morphologically Complex Turkish Words" (Bilgin, 2016).

depends on the comprehensiveness of the lexical database that is used in the analysis. Otherwise, words which are not proper nouns may be identified as proper noun. Despite a comprehensive and well-defined database, complications may still arise. For instance, a common noun, or an adjective, or a number may be used as a proper noun in a text—e.g., "Five Armies" (Tolkien 2009).

Figure 7. Identification process of proper nouns (process diagram illustrating how a word is identified as a proper noun by the tool)[6]



[6] The step "analyse morphologically" requires further development in the software.

The results of the process described above are presented in table 8. As there are at least 278 results to be listed, only the first 15 entries are indicated in the table due to lack of space.

Table 8. List of proper nouns (first 15 entries)

Source Text	**Target Text 1**	**Target Text 2**
Baggins	Bilbo	Bilbo
Smaug	Thorin	Thorin
Gollum	Gandalf	Gandalf
Balin	Baggins	Baggins
Beorn	Smaug	Gollum
Bombur	Gollum	Smaug
Mirkwood	Balin	Beorn
Elrond	Beorn	Balin
Elvenking	Bombur	Bombur
Gloin	Bay Baggins	Bay Baggins
Bofur	William	Kuyutorman
Wargs	Kasvetormanı	William
Dwalin	Elrond	Goblinler
Arkenstone	Goblinler	Elrond
Bilbo Baggins	Goblin	Goblin

5.2.2 Non-defined Words. Tolkien (2009) coins new words in the novel, and Duymaz (2020) analyzes their translation. In order to carry out the same analysis, a similar process as in 'proper nouns' is followed by the tool. Basically, words that are not included in the lexical database are identified as non-defined words. The drawback of this process is that it retrieves both coined words and other words such as deliberately or unintentionally mistyped ones. The researcher should eliminate the irrelevant ones to obtain neologisms. Except from that, the results match with the ones in Duymaz's research. In table 9, total numbers of non-defined words and first 15 of them in the list are presented.

Table 9. Non-defined words (first 15 entries)

Source Text	**Target Text 1**	**Target Text 2**
hobbit	dwarf'ın	hobbit

Source Text	Target Text 1	Target Text 2
ork	dwarfs	elfleri
exceptfor	jöleymişçesine	orkların
hobbitlike	dwarwes	hobbitler
smokering	dwarwish	köselemsi
morninged	goblin	goblinlerle
bewuthered	hobgoblin	ding
bebother	hobbitlerin	yiyemedüh
dwarvish	orc	warg
tomorrer	ork	didi
manflesh	runlarla	konuşuyon
thinkin	macerasal	geçiyi
runnin	plop	sökiyi
burrahobbit	runlarda	rünü
nassty	güsel	hobgoblinler
74 entries	**347 entries**	**716 entries**

The substantial difference between the total number of entries in Target Text 2 and the others can be justified when all the entries of Target Text 1 and Target Text 2 are compared manually. Most of the entries in Target Text 2 consist of the words that are mistyped intentionally to express accents of the characters.

5.2.3 Exclamations. Results in this part are retrieved only if there is an exclamation mark at the end of a phrase and it has been used more than once in the text. Table 10 shows the entries defined as exclamations.

Table 10. Exclamations

Source Text	Count	Target Text 1	Count	Target Text 2	Count
O	7	Aman Tanrım	2	Teşekkür ederim	2
ha	3	Hey	5	Hey	6
Snap	2	ha	3	ha	3
my lad	2	Devam edelim	2	Ez, parçala	2
curse it	2	kahretssin	2	Lanet olssun ona	3
Attercop	2	Kafadanbacaklı	2	Eklembacak	2

Source Text	Count	Target Text 1	Count	Target Text 2	Count
Attercop	2	Kafadanbacaklı	2	Eklembacak	2
and South away	2	ve Güneye doğru	2	Git güneye	2
To arms	2	Silahlanın	2	Silahlanın	2
To me	2	Bana gelin	2	Bana gelin	2
		Kartallar	2	Kartallar	2
		Tra - la - la - lalle	2		
		Gelin hadi gerisin geri vadiye	2		

5.2.4 Measurement Units. In this part, measurement units are presented along with the words that come immediately before and after them, e.g., "many **miles** round." The software is taught the measurement units, and the resulted entries are provided in table 11. This category can be significant as the researcher can analyze them and determine whether they conform to the source norms or target norms.

Table 11. Measurement units (first 15 entries)

Source Text	Target Text 1	Target Text 2
many **miles** round	bir **mil** öteden	bir **mil** öteden
a **mile** off	yüzlerce **mil** uçup	yüz **metre** uçtuktan
small Five **feet** high	Kapı beş **ayak** yüksekliğinde	buçuk **metre** yüksekliğinde
a **mile** or	iki **mil** daha	bir **mil** daha
passed two **feet** off	yüzlerce **mil** uzakta	bir **metre** ötemizden
hundred **miles** away	kapı beş **ayak** yüksekliğinde	yüzlerce **mil** uzakta
say five **feet** high	[Ev']i **millerce** geride	buçuk **metre** yüksekliğinde
House **miles** behind	dek dört **ayak** üzerinde	bir **metre** yanından
gongsPound **pound** far	ayak üç **ayak** yanında	iki **metre** havaya
a **yard** to	oturur dört **ayak** da	kavuşmadan **millerce** uzakta
jumped seven **feet** forward	üzerinden üç **ayak** havaya	sabahıydı **Millerce** ilerleyip
be **miles** on	ve yedi **ayak** ileriye	üç **metre** havaya
gone **miles** and	zamanında dört **ayak** üzerinde	bir **mil** aşağısındaki
air ten **feet** and	basmadan **millerce** yol	tarafından **millerce** yaylaya

Source Text	**Target Text 1**	**Target Text 2**
a **mile** below	sabahıydı **Millerce** yol	birkaç **mil** kuzeyindesiniz
36 entries	**40 entries**	**36 entries**

5.2.5 Personal Titles. Personal titles are taught to the tool, and the results are obtained as given in table 12.

Table 12. Personal titles (first 15 entries)

Source Text	**Target Text 1**	**Target Text 2**
became **Mrs**. Bungo	Took, **Bayan** Bungo	Gerçi **Bayan** Bungo
our **Mr.** Baggins	bizim **Bay** Baggins	bizim **Bay** Baggins
dear **sir**! Let	ederim, **sayın bayım**	biliyorum, **Bay** Bilbo
name, **Mr.** Bilbo	biliyorum, **Bay** Bilbo	kadar **Bay** Baggins
that **Mr.** Baggins	gibi **Bay** Baggins	düşündü **Bay** Baggins
thought **Mr.** Baggins	ederim **sayın** efendim	zavallı **Bay** Baggins
poor **Mr.** Baggins	düşündü **Bay** Baggins	olan **Bay** Baggins
thought **Mr.** Baggins	zavallı **Bay** Bilbo	Çıkmazlı **Bay** Baggins
plain **Mr.** Baggins	başlayan **Bay** Baggins	ve **Bay** Baggins
and **Mr.** Baggins	yavan **Bay** Baggins	zavallı **Bay** Baggins
estimable **Mr.** Baggins	ve **Bay** Baggins	ediyorum **Saygıdeğer** Bay
Then **Mr.** Baggins	durumu **Saygıdeğer** Bay	Derken **Bay** Baggins
chose **Mr.** Baggins	Sonra **Bay** Baggins	de **Bay** Baggins
chosen **Mr.** Baggins	de **Bay** Baggins	Ben **Bay** Baggins
poor **Mr.** Baggins	Ben **Bay** Baggins	zavallı **Bay** Baggins
91 entries	**79 entries**	**71 entries**

5.2.6 Most Frequent Words. For this category, the stop words are first removed. ‘Stop words’ mean commonly used words in a language such as the subject ‘I’ and determiners and conjunctions. In this way, they are prevented from affecting the results. Finally, most frequent words are obtained as shown in table 13.

Table 13. Most frequent words (first 15 entries)

Source Text	**Count**	**Target Text 1**	**Count**	**Target Text 2**	**Count**
Bilbo	443	Bilbo	349	Bilbo	338
dwarves	273	iyi	188	uzun	201
long	214	uzun	183	büyük	196
Thorin	203	dek	161	zaman	193
back	199	büyük	156	iyi	178
great	197	doğru	145	doğru	149
time	195	Thorin	144	Thorin	148
good	157	geri	132	küçük	118
Gandalf	149	küçük	128	Gandalf	117
goblins	138	yalnızca	122	son	117
dark	133	cüceler	113	Bilbo'nun	101
made	126	Gandalf	112	süre	101
mountain	120	zaman	103	cüceler	101
thought	120	üzerine	103	hâlâ	95
hobbit	113	aşağı	100	tek	92

'Most frequent words' are not included in Duymaz's research, yet they are added in this research because they can lead the researcher to further questions and different interpretations. For example, the researcher can make inferences regarding the theme of the texts and whether they match in their translations. In the case of *The Hobbit*, we can see that words regarding nature and the words like 'back' and 'end' may imply the journey of the main character Bilbo Baggins. Also, the researcher may want to explore the preference for using specific lexical items where other options may be equally available in the language.

6. Discussion

The initial objective of this project was to facilitate exhaustive analysis to extract operational norms. The reason behind limiting it to 'operational norms' is that the techniques or functions in text mining are particularly suited to identify operational norms. Although it was not Toury's aim to "point to strict statistical methods for dealing with translational norms" or "supply sampling rules," the findings of this research affirm Toury's perspective on the possibility of exhaustive analysis on "many more and much bigger samples" with the help of advances in computing world (2012, 91–92). The findings indicate that text mining can make it possible and easier to analyze these specific types of norms exhaustively, hence in a way that cannot be done by sole human effort, especially on large corpora. However, it should be noted that the current analysis has not been able to cover all the categories under 'textual-linguistic norms,' especially the ones that have cultural elements or expressive language. On the other hand, this should not deter the researcher from developing the tool as the advances in NLP can make it possible. The project thus has the potential to contribute to the development of descriptive branch of TS.

7. Conclusion

The present research was conducted to facilitate exhaustive analysis on operational norms as introduced by Toury since it is not possible due to human limitation. Drawing on Toury's study, the findings are presented under two norms: 'matricial norms' and 'textual-linguistic norms.' Also, the categories analyzed in Duymaz's research are included in this project to test the effectiveness of the tool in relation to analysis conducted by a human. The results obtained under the first type of norms revealed the statistics regarding the distribution of linguistic material in the texts, whereas the second one showed the selection of linguistic material. The findings suggest that the aim of this research was fulfilled as the tool provided a considerable number of quantitative data and made further insights possible.

It should be noted that the findings are subject to three limitations. The first limitation is regarding the preparation of data for analysis. The efficacy of the tool is limited to the image quality of the texts and the pre-editing process if a scanned document is analyzed. Secondly, it is important to bear in mind that although the tool can process as many texts as a researcher wishes to analyze, the time required to process them depends on the capacity of the server. Finally, further tests are needed to ensure the tool's viability even though it is technically suitable for all types of texts.

Despite all limitations of the proposed tool, it can be said that the expected outcome is achieved in this project. This work contributes to descriptive translation analysis by offering the possibility to analyze on larger quantities of data and overcome human limitations to a certain extent. What is now needed is further research and effort to improve the tool as it has the potential for being a worthwhile contribution to DTS.

References

Baker, Mona. 1993. “Corpus Linguistics and Translation Studies: Implications and Applications.” In *Text and Technology: In Honour of John Sinclair*, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–250. Philadelphia: John Benjamins.

———. 2000. “Towards a Methodology for Investigating the Style of a Literary Translator.” *Target* 12 (2): 241–266. doi:10.1075/target.12.2.04bak.

Bilgin, Orhan. 2016. “Frequency Effects in the Processing of Morphologically Complex Turkish Words.” Master’s thesis, Boğaziçi University.

Cheesman, Tom, Kevin Flanagan, Stephan Thiel, Jan Rybicki, Robert S. Laramee, Jonathan Hope, and Avraham Roos. 2017. “Multi-Retranslation Corpora: Visibility, Variation, Value, and Virtue.” *Digital Scholarship in the Humanities* 32 (4): 739–760. doi:10.1093/llc/fqw027.

Çalışkan, Sevilay. 2020. “Text Mining Analysis of Translation, Social Communication and Literary Writing for Turkish.” Master’s thesis, Bilkent University.

Duymaz, Zeynep. 2020. “John Ronald Reuel Tolkien’in *The Hobbit* Adlı Eserinin Türkçe Çevirilerinin Yeniden Çeviri Varsayımı ve Çocuk Edebiyatı Kapsamında İncelenmesi.” [An analysis of the Turkish translations of John Ronald Reuel Tolkien’s *The Hobbit* within the framework of retranslation hypothesis and children’s literature.] Master’s thesis, Istanbul Okan University.

Jackson, Peter, and Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization.* Amsterdam: John Benjamins.

Kit, Chunyu, and Jian-Yun Nie. 2015. “Information Retrieval and Text Mining.” In *The Routledge Encyclopedia of Translation Technology*, edited by Chan Sin-wai, 494–535. London: Routledge.

Li, Defeng. 2017. “Translator Style: A Corpus-Assisted Approach.” In *Corpus Methodologies Explained: An Empirical Approach to Translation Studies*, edited by Meng Ji, Michael Oakes, Defeng Li, and Lidun Hareide, 103–136. Abingdon, Oxon: Routledge.

Monti, Johanna, Annibale Elia, Alberto Postiglione, Mario Monteleone, and Federica Marano. 2011. “In Search of Knowledge: Text Mining Dedicated to Technical Translation.” In *Proceedings of ASLIB 2011 Translating and the Computer Conference*.

Oi Yee, Olivia Kwong. 2015. “Natural Language Processing.” In *The Routledge Encyclopedia of Translation Technology*, edited by Chan Sin-wai, 563–577. London: Routledge.

Patton, Jon M., and Fazlı Can. 2012. “Determining Translation Invariant Characteristics of James Joyce’s *Dubliners*.” In *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, edited by Michael P. Oakes and Meng Ji, 209–229. Amsterdam: John Benjamins.

Tolkien, John Ronald Reuel. 1996. *Hobbit: Oradaydık ve Şimdi Buradayız*. Translated by Emel İzmirli. Istanbul: Altıkırkbeş / Mitos.

———. 2009. *The Hobbit*. London: HarperCollins.

———. 2015. *Hobbit*. Translated by Gamze Sarı. Istanbul: İthaki.

Toury, Gideon. 1978. “The Nature and Role of Norms in Translation.” In *Literature and Translation: New Perspectives in Literary Studies; With a Basic Bibliography of Books on Translation Studies*, edited by James S. Holmes, José Lambert, and Raymond van den Broeck, 83–100. Leuven: Acco.

———. 1995. *Descriptive Translation Studies — and Beyond.* Amsterdam: John Benjamins.

———. 2012. *Descriptive Translation Studies — and Beyond*. 2nd expanded ed. Amsterdam: John Benjamins.

Venuti, Lawrence. 1995. *The Translator’s Invisibility*. New York: Routledge.

Zanini, Nadir, and Vikas Dhawan. 2015. “Text Mining: An Introduction to Theory and Some Applications.” *Research Matters: A Cambridge Assessment Publication*, no. 19, 38–44. https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf.