

TURKISH - ENGLISH MACHINE TRANSLATION SYSTEM

by

SümeYra Yılmaz & İrem Uğuz

Submitted to the Department of Computer
Engineering in partial fulfillment of
the requirements for the degree of
Bachelor of Science

Undergraduate Program in Computer Engineering
Boğaziçi University
Spring 2020

TURKISH - ENGLISH MACHINE TRANSLATION SYSTEM

APPROVED BY:

Prof. Tunga Güngör
(Project Supervisor)

DATE OF APPROVAL: 06.04.2020

ABSTRACT

TURKISH - ENGLISH MACHINE TRANSLATION SYSTEM

In this project, a Turkish-English parallel corpus is created and a machine translation system based on that corpus is constructed. We started by analyzing an already existing corpus. Then, we decided to organize the existing corpus so that the new version would be more reliable to use. After organizing the already existing corpus [1], we created a statistical machine translation system using Moses [2]. Experiments with different sizes of the corpus are conducted on the system to optimize the result. We also tried to create a neural machine translation system, however we didn't have enough computer resource to run any neural machine translation tool. Therefore, we decided to conduct experiments with processed input to find better results.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF ACRONYMS/ABBREVIATIONS	vi
1. INTRODUCTION AND MOTIVATION	1
2. STATE OF THE ART	2
2.1. Turkish-English Parallel Corpus	2
2.2. SMT and NMT	3
3. METHODS	4
3.1. BOUN Turkish-English Parallel Corpus	4
3.1.1. The reference links for the texts	6
3.2. Installation of Moses and Giza	11
3.3. Running Moses	14
3.3.1. Corpus Preparation	15
3.3.2. Language Model Training	16
3.3.3. Training the Translation System	16
3.3.4. Tuning	16
3.3.5. Testing	17
3.4. Data Preprocessing	18
3.4.1. Stemming words	19
3.4.2. Removing Punctuation	20
3.4.3. Stopword Removal	21
3.4.4. Shuffling Data	22
3.4.5. Changing Train Data	23
3.4.6. 4-gram Language Model	23
3.5. NMT	24
4. RESULTS	25
4.1. Turkish-English Parallel Corpus	25
4.2. Statistical Machine Translation System	29
4.2.1. Corpus With Different Sizes And Pre-Processing Techniques . .	30
4.2.2. BLEU Scores Of Different Trials	31

4.2.3. Example Translations With Non-shuffled Corpus	31
4.2.4. Example Translations With Shuffled Corpus	34
5. CONCLUSION AND DISCUSSION	36
6. FUTURE WORK	39
REFERENCES	40

LIST OF ACRONYMS/ABBREVIATIONS

XCES	Corpus Encoding Standard for XML
XML	Extensible Markup Language
SMT	Statistical Machine Translation
NMT	Neural Machine Translation

1. INTRODUCTION AND MOTIVATION

Firstly, we decided to create a Turkish-English parallel corpus that can be later used with machine translation applications. A corpus means a large collection of text produced by real users. A parallel corpus is a collection of two corpora, in a way that one is the translation of the other, and texts are aligned, which are Turkish and English in this project. To create a more comprehensive corpus than the already existing ones, we started with checking the status of these corpora and collecting translated data. We also did a literature scan to learn sentence alignment algorithms as we were planning to create our corpus.

After checking the state of the art, we decided that the existing BOUN corpus [1] is a very extensive and we could organize this to create a more organized one to use with machine translation systems. We divided the corpus into separate texts that contain different types of books. We thought this classification would give better results for different contexts.

Then we moved on to the machine translation systems. We planned to work with statistical and neural machine translation in that order. For SMT, we used Moses, a statistical machine translation system, and Giza, a word-aligning tool can be used with Moses. Moses [3] is a statistical machine translation system that allows you to automatically train translation models for any language pair. Moses offers two types of translation models: phrase-based and tree-based. We used phrase-based translation model. We used both the corpus we designed and the pre-existing corpus [1]. Details of how we organized the corpus and used Moses and Giza can be found in the Methods section.

Unfortunately, because of the technical problems like our computer power, we would not be able to move on to neural machine translation, instead we worked on our data to get better results with Moses.

2. STATE OF THE ART

2.1. Turkish-English Parallel Corpus

There have been several English-Turkish parallel corpora which are done prior to our project. We have inspected those corpora to see how can we improve the current sources.

There has been some work on creating corpus from religious texts. [4] is a very extensive corpus that contains tokenized and aligned Bible in XML format for more than 10 languages. [5] also has an aligned Bible corpus for Turkish and English. The Bible corpus contains nearly 93k lines. Also, [1] contains a religious book about Quran which is also aligned.

[6] is an aligned corpus that contains news articles in Turkish, English and Kurdish. The corpus is aligned in XCES format which is very commonly used format in parallel corpus research. The size of the corpus is very nice as well, containing 35080 sentences.

[5] contains aligned Bible (as mentioned above), user manuals and short stories. The corpus is encoded in XML format and is aligned. The manual data is very extensive data which nearly contains 150000 lines. However, the corpus also is a little unclean and contains unnecessary tag words in the XML file.

Another corpus is [7] that contains mainly book 1984 and a short TDK manual. The fiction book 1984 is not very well aligned but it is a clean translation. The TDK manual is aligned but very short, containing only 10 lines.

[8] is also an available Turkish-English parallel corpus which is aligned and ready for use. However, the corpus data is not open for examination and is only usable in projects.

The main corpus that we are examining is [1] which is a very large (nearly 600000 lines) aligned Turkish-English corpus. The corpus contains several types of books like religious books, theater play, fiction and web-site guide. However, the corpus contains all the data together in one place and needs to be divided into smaller sections for the ease of use.

2.2. SMT and NMT

After organizing our corpus, we started to work on creating a SMT system using Moses. Thus we did a research about how Moses can be used with Turkish-English parallel corpus. [9] is a SMT system made with Moses by Tubitak, and has high BLEU scores(around 50), which is our base criterion. In this research [9] they experimented with hierarchical phrasebased models.

We also read the thesis of our seniors that has also worked with Moses on a different project. They helped us to understand how to install and run Moses.

3. METHODS

3.1. BOUN Turkish-English Parallel Corpus

There exists a Turkish English parallel corpus created by the contributions of Bogazici University and ITU [1]. It has two aligned files, one in English and one in Turkish and composed of many texts. It contains official documents, movie scenarios, and religious and literary texts.

We want these aligned files separately to use it in a machine translation system and to create a well-organized corpus. Also, since the alignment degree of the corpus is unknown, we decided to create a better-aligned corpus using these.

Our first attempt is to divide this huge text files into smaller meaningful parts. The corpus was created in this way, texts are divided into different sized parts, which are combined in mixed order, and some of them are repeated. Hence, we started to divide these files.

We worked synchronously on the same text. The size of the whole corpus is nearly 700000 lines, so we divided it into two sections. One of us would work on the first 350000 lines of the text where the other one would work on the last 350000 lines of the corpus. That way, tow people could work on the text synchronously. We put English and Turkish files to separate links. The Drive links can be found below.

To separate one text file from the corpus at first we searched for the sentences in Google and found the book that the sentence belongs to. Then continue looking through the text to see the end of the text piece. Around the end of the text piece, we check each line to see if they belong to the same book. After finding the end and start line of the text in one language, the same lines of the other language corpus is checked and that part is selected as well.

After finding the list of the books that are used in the previous research, it became easier to find which book that the sentence belongs to. We could check if the book contains the given sentences from the book's PDF file.

After determining the end and the start lines of the text piece, we would put it to a separate text file and upload this file to the Drive. Finding the same text in another language is easy because the corpus is sentence aligned. By checking corresponding line numbers, we would also be checking if the alignment is correct. The names of the books in the corpus can be seen in Figure 3.1.1.

BOOKS										
ID	NAME	CATEGORY	TYPE	SUB-TYPE	# OF WORDS	LENGTH	FORMAT	QUALITY	SOURCE	ADDITIONAL INFORMATION
B001	harry potter-felsefe tagı	kitap	roman	fantastik	56.000	170 sayfa	txt,doc,pdf	good	http://www.alkitap.com/arsiv.asp	1 pdf EN, pdf+txt+doc TR
B002	harry potter-sirilar odasi	kitap	roman	fantastik	67.000	189 sayfa	txt,doc,pdf	good	http://www.alkitap.com/arsiv.asp	1 pdf EN, pdf+txt+doc TR
B003	harry potter-azkaban tutsağı	kitap	roman	fantastik	84.800	178 sayfa	txt,doc,pdf	good	http://www.alkitap.com/arsiv.asp	1 pdf EN, pdf+txt+doc TR
B004	harry potter-ateş kadehi	kitap	roman	fantastik	150.000	302 sayfa	txt,doc,pdf	good	http://www.alkitap.com/arsiv.asp	1 pdf EN, pdf+txt+doc TR
B005	harry potter-sünmürcükenin yoldaşlığı	kitap	roman	fantastik	200.000	418 sayfa	doc, pdf	good	http://www.alkitap.com/arsiv.asp	1 pdf EN, pdf+doc TR-sayfa nolari var
B006	yüzkülinin efendisi-yüzkü kardeşliği	kitap	roman	fantastik	142.000	450 sayfa	txt,doc,pdf, lit	good	http://www.e-kutuphane.net/	1 pdf EN, txt+doc+lit TR
B007	yüzkülinin efendisi-iki kule	kitap	roman	fantastik	119.000	380 sayfa	txt,doc,pdf, lit	good	http://www.e-kutuphane.net/	1 pdf EN, txt+doc+lit TR-bazı yerde bölümler yarım
B008	yüzkülinin efendisi-krallın dönüşü	kitap	roman	fantastik	106.000	310 sayfa	txt,doc,pdf, lit	good	http://www.e-kutuphane.net/	1 pdf EN, 1 txt TR
B009	1984 by George Orwell	kitap	roman	bilimkurgu	65.000	220 sayfa	txt	good	http://ekitap.kolayweb.com/	1 pdf EN, html+txt TR ----DIALOG SEKÜNDE..
B010	Macbeth by Shakespeare	kitap	oyun	trajedi	18.200	32 sayfa	txt, html, pdf	adequate	http://ekitap.kolayweb.com/	1 txt EN, 1 txt+pdf TR
B011	Red smeatary by Stephen King	kitap	roman	korku	87.000	142 sayfa	txt, pdf	good	http://www.kitap.persil.com/	1 pdf EN, 1 pdf TR
B012	Da Vinci Şifresi	kitap	roman	polisye	77.200	295 sayfa	pdf	good	http://www.kitap.persil.com/	1 lit EN, 2 txt TR
B013	Descartes- Discourse on method	kitap	felsefe		24.700	47 sayfa	lit, txt	good	http://www.e-kutuphane.net/	1 lit EN, 2 txt TR
B014	Bacon - New Atlantis	kitap	felsefe	politik	13.000	33 sayfa	lit, txt	good	http://www.e-kutuphane.net/	1 lit EN, 2 txt TR
B015	Plato - Statesman	kitap	felsefe	politik	18.700	106 sayfa	lit, txt	good	http://www.kitap.persil.com/	1 lit EN, 2 txt TR ----diyalog seklinde..
B016	Tommaso Campanelli - City of Sun	kitap	felsefe	politik	23.700	40 sayfa	lit, txt	good	http://www.kitap.persil.com/	1 lit EN, 2 txt TR felsefe
B017	Dostoyevski - Verahından Notlar	kitap	roman	felsefik	30.000	98 sayfa	lit, txt	good	www.freeELiterature.com	1 lit EN, 1 txt TR
B018	Thoreau- Hakikat Yonetime Karşı	kitap	felsefe	politik	8.300	21 sayfa	lit, doc	good	www.freeELiterature.com	1 lit EN, 1 doc TR
B019	Tolstoy - Anna karanina	kitap	roman	dram	351.000	883 sayfa	lit	good	www.freeELiterature.com	1 lit EN, 2 lit TR(cilt1-cilt2 seklinde)
B020	Aristoteles - Atinaların Devleti	kitap	felsefe	politik	24.400	43 sayfa	doc, txt	good	http://www.kitap.persil.com/	1 doc EN, 1 txt TR
B021	Plato - Republic	kitap	felsefe	politik	43.400	349 sayfa	html, lit, txt	good	www.freeELiterature.com	1 lit EN, 1 txt+1 doc TR
B022	Mark Twain - Tom Sawyer	kitap	roman	macera	71.000	139 sayfa	doc, lit	very good	www.insanizm.com	1 lit EN, 1 doc TR
B023	Voltaire - Candide	kitap	roman	dram	36.600	80 sayfa	doc, txt	good	http://www.kitap.persil.com/	1 doc EN, 1 txt TR
B024	Clauewitz - Savaş	kitap	inceleme	savas	66.000	105 sayfa	doc, html	adequate	http://www.kitap.persil.com/	1 html EN, 4 cilt olarak doc TR
B025	Lenin - Devlet ve İhtilal	kitap	inceleme		28.600	90 sayfa	lit, doc	good	www.insanizm.com	1 lit TR, 7 doc TR(herbiri 1 chapter)
B026	Plato - Apology	kitap	soşyal	politik	11.600	42 sayfa	lit, html, txt	good	http://www.kitap.persil.com/	1 lit TR, 1 txt+1 html EN
B027	Cicero - yasilik ve Dostluk	kitap	felsefe		22.000	65 sayfa	lit, html	good	http://www.kitap.persil.com/	1 lit EN, 1 html TR
B028	Stephen King - Green Mile	kitap	roman	duygusal	194.000	443 sayfa	lit, pdf, txt	good	www.freeELiterature.com	1 lit TR, 1 txt+1 pdf EN
B029	Carus - On the nature of Things	kitap	felsefe		74.000	175 sayfa	doc,txt	adequate	www.freeELiterature.com	1 doc EN, 1 txt TR ----şir seklinde yazılmış
B030	Tolstoy - Master and Man	kitap	roman	dram	19.200	64 sayfa	doc, txt	good	http://www.kitap.persil.com/	1 pdf TR, 1 txt EN
B031	Tolstoy - Ivan ilic	kitap	roman	dram	15.800	32 sayfa	doc, txt	good	http://www.kitap.persil.com/	1 doc TR, 1 txt EN
B032	belgarlad-ikehanetin oyuncagi	kitap	roman	fantastik	79.540	157 sayfa	htm,pdf	adequate	http://www.gutenberq.org/	1 pdf TR, 1 htm EN
B033	belgarlad-2buyuculer kralicesi	kitap	roman	fantastik	106.000	195 sayfa	htm,pdf	adequate	http://www.gutenberq.org/	1 pdf TR, 1 htm EN
B034	belgarlad-şahinbazın tuzağı	kitap	roman	fantastik	97.000	180 sayfa	htm,pdf	adequate	http://www.gutenberq.org/	1 pdf TR, 1 htm EN
B035	belgarlad-4buyulu sato	kitap	roman	fantastik	120.000	206 sayfa	htm,pdf	adequate	http://www.gutenberq.org/	1 pdf TR, 1 htm EN
B036	belgarlad-seluncunun son oyunu	kitap	roman	fantastik	116.560	197 sayfa	htm,pdf	adequate	http://www.gutenberq.org/	1 pdf TR, 1 htm EN
B037	Arthur Clarke-2001A Space Odyssey	kitap	roman	bilimkurgu	61.850	138 sayfa	doc,pdf	good	http://www.gutenberq.org/	1 doc TR, 1 pdf+1 doc EN
B038	Arthur Clarke- c_Clarke-Rama_2	kitap	roman	bilimkurgu	114.470	245 sayfa	txt,txt	good	http://www.gutenberq.org/	1 txt TR, 1 txt EN
B039	Arthur Clarke- rendezvous with rama	kitap	roman	bilimkurgu	72.000	193 sayfa	doc,doc	good	http://www.gutenberq.org/	1 doc TR, 1 doc EN
B040	bernard shaw - Sesar ve Kleopatira	kitap	oyun	drama	39.000	102 sayfa	html, txt	good	http://www.kitap.persil.com/	1 txt EN, 1 html TR
B041	kafka - Metamorphosis	kitap	hikaye		15.700	28 sayfa	doc,txt	adequate	www.insanizm.com	1 doc EN, 1 txt TR
B042	goethe - faust	kitap	gırlı		12.700	40 sayfa	doc,txt	adequate	www.insanizm.com	1 doc EN, 1 txt TR
B043	göçli- Teras Buluş	kitap	roman	kurgu	51.760	94 sayfa	pdf,txt	good	http://www.kitap.persil.com/	1 pdf TR, 1 txt EN
B044	Eleanor_H_Porter-Pollyanna	kitap	roman	iyimserlik	95.000	301 sayfa	txt,txt	very good	http://www.gutenberq.org/	1 txt TR, 1 txt EN
B045	Anatole France - Tihals	kitap	roman	ask macera	36.000	69 sayfa	txt,doc	good	http://www.kitap.persil.com/	1 txt EN, 1 doc TR
B046	Dostoevsky - karamazov kardeşler	kitap	roman	dram	350.000	562 sayfa	txt,doc	good	http://www.kitap.persil.com/	1 txt EN, 2 doc TR(iki cilt seklinde)
B047	Turgenev - rudin	kitap	roman		53.460	118 sayfa	txt,lit	good	http://www.gutenberq.org/	1 txt EN, 1 lit TR
B048	Stevenson - Markheim	kitap	hikaye		5.000	11 sayfa	txt,doc	good	http://www.gutenberq.org/	1 txt EN, 1 doc TR

Figure 3.1.1 - Books in the corpus

B049	Dostoyevski-KUMARBAZ	kitap	roman	dram	62 880	126 sayfa	txt,lt	good	http://www.gutenberg.org/	1 txt EN, 1 lt TR
B050	Goethe - Iphigenia in Tauris	kitap	oyun	drama	19 830	45 sayfa	txt,lt	very good	http://www.gutenberg.org/	1 txt EN, 1 lt TR
B051	Lermontov - A Hero of Our Time	kitap	roman	dram, macera	37 000	68 sayfa	txt,doc	good	http://www.gutenberg.org/	1 txt EN, 1 doc TR
B052	Moliere - The Imaginary Invalid	kitap	oyun	eleştiri	14 900	61 sayfa	txt,doc	adequate	http://www.gutenberg.org/	1 txt EN, 1 doc TR
B053	G. Leroux - Mystery of Yellow Room	kitap	roman	polisiye	47 250	85 sayfa	txt,doc	good	www.freeELiterature.com	1 txt EN, 1 doc TR
B054	Jack London - The Call of the Wild	kitap	roman		33 800	63 sayfa	txt,lt	good	www.freeELiterature.com	1 txt EN, 1 lt TR
B055	Dostoyevski - Devils	kitap	roman	siyasal	260 000	440 sayfa	html,lt	adequate	http://www.kitap.parisi.com/	1 html EN, 1 lt TR (turkce karakterler bozuk...)
B056	Baltac - Eugenie Grandet	kitap	roman		55 750	93 sayfa	txt,doc	good	www.freeELiterature.com	1 txt EN, 1 doc TR
B057	Baltac - Hidden Masterpiece	kitap	hikaye		13 300	27 sayfa	txt,lt	good	www.insanizm.com	1 txt EN, 1 lt TR
B058	Anatole France - Penguin Island	kitap	roman		52 800	91 sayfa	txt,doc	very good	http://www.kitap.parisi.com/	1 txt EN, 1 doc TR
B059	Chamisso - Peter Schlemihl	kitap	roman	psikoloji	38 380	75 sayfa	txt,lt	good	http://www.kitap.parisi.com/	1 txt EN, 1 lt TR
B060	Oscar Wilde-The Happy Prince,Tales	kitap	hikaye	cocuk	10 700	18 sayfa	txt,doc	very good	www.insanizm.com	1 txt EN, 1 doc TR
B061	Dostoyevsky - Crime and Punishment	kitap	roman	psikoloji	203 000	330 sayfa	txt,doc,lt,pdf	good	http://www.kitap.parisi.com/	1 txt+pdf+doc EN, 2 lt+2 pdf TR (2 cilt halinde...)
B062	20 Soruda Evrim Teorisinin Çöküşü	kitap	dini		20514	66	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B063	Akılsız Kur'anlı Nasıl Yorumlar	kitap	dini		19143	56	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B064	Allah Akıla Bilir	kitap	dini		41658	112	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B065	Allah İpin Yaşamak	kitap	dini		20379	59	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B066	Allah'ın İsimleri	kitap	dini		48301	145	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B067	Allah'ın Munia Sanatı	kitap	dini		13567	38	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B068	Renk Sanatı	kitap	dini		16382	46	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B069	Allah Korkusu	kitap	dini		22758	69	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B070	Allah Sevgisi	kitap	dini		19180	56	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B071	Akıllıcağ	kitap	dini		16794	53	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B072	Amerikan Ulusal Bilimler Akademisi	kitap	dini		42362	115	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B073	Atom Mucizesi	kitap	dini		24161	60	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B074	Bilgilerdeki Yaratılış Mucizesi	kitap	dini		26090	72	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B075	Biyomimetik Teknoloji Doğayı Taklit E	kitap	dini		27951	85	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B076	Cahiliye Toplumunda İnsan Karakteri	kitap	dini		27888	72	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B077	Canlılardaki Fedakarlık Ve Akılcı Davi	kitap	dini		29651	78	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B078	Çocuklar Sizin İpin 1	kitap	dini		17535	50	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B079	Çocuklar Sizin İpin 2	kitap	dini		16215	46	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B080	Çözüm Kur'an Ahlakı	kitap	dini		32231	87	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B081	Danwinizm'in Kanlı İdeolojisi: Fağzım	kitap	dini		37289	77	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B082	Danwinizm'in Karanlık Büyüsü	kitap	dini		32863	61	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B083	Deccal'in Ayını Terör	kitap	dini		23418	67	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B084	Çocuklar Danwin Yalan Söyledi	kitap	dini		15212	53	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN
B085	Danwinizm'in İnsanlığa Getirdiği Belal	kitap	dini		30469	77	doc,rtf	good	www.harunyahya.org	1 doc TR 1 doc EN

SHORT TEXTS

ID	NAME			# OF WORDS	LENGTH	FORMAT	QUALITY	SOURCE	ADDITIONAL INFORMATION
T001	bilgent yönetmelik	kısa metin	yönetmelik	2 800	7 sayfa	.doc	adequate	www.bilgent.edu.tr	atlamalar, farklı cümle yapıları var.
T002	erhan sigorta	kısa metin	poliçe	3 300	9 sayfa	.doc	very good	ceviribilim bolumu	başlı doldurulacak form sonrası normal text
T003	ileri eng bou	kısa metin	tanıtım	2 440	7 sayfa	.doc	adequate	www.boun.edu.tr	atlamalar, farklı cümle yapıları
T004	record 2006	kısa metin	mail	345	1 sayfa	.doc	good	mail	informal, kısa
T005	web örnek	kısa metin		499	2 sayfa	.doc	very good	tercuman bürosu	formal, kalitesi yüksek
T006	working capital	kısa metin	rehber	3 200	10 sayfa	.doc	very good	tercuman bürosu	işletme sermayesi rehber el kitabı
T007	yorum	kısa metin	kısa yazısı	61 880	125 sayfa	.doc	adequate	hurriyet.com, zaman.com	9'a adet yorum, daha artacak.
T008	hotel/bakkal	kısa metin	reklam	432	2 sayfa	.doc	adequate	http://www.kitap.parisi.com/	otel reklam afişi
T009	marş	kısa metin	şir	101	1 sayfa	.doc	adequate		istiklal marşı çevirisi
T010	ninni	kısa metin	hikaye	485	2 sayfa	.doc	good	http://www.kitap.parisi.com/	hikaye
T011	seftali	kısa metin	hikaye	358	2 sayfa	.doc	good	http://www.kitap.parisi.com/	hikaye
T012	yazlıkaya	kısa metin	tanıtım	1	sayfa	.doc	good	http://www.kitap.parisi.com/	düzyazı çeviri
T013	savaş dansları	kısa metin	şir	1	sayfa	.doc	adequate		şir
T014	sen suçu değilsin	kısa metin	şir	1	sayfa	.doc	adequate		şir
T015	sevin birbirinizi	kısa metin	şir	1	sayfa	.doc	adequate		şir

Figure 3.1.1 - Continued

T016	unutma	kısa metin	şir	1 sayfa	.doc	adequate	şir
T017	varoluş ucu	kısa metin	şir	1 sayfa	.doc	adequate	şir

OTHER RESOURCES

ID	NAME	QUANTITY	FORMAT	QUALITY	SOURCE	ADDITIONAL INFORMATION
X001	el yazıları	6 adet	rtf, txt, sub	adequate	www.cinsalhyazi.com	süre sınırlıdır, alırken gerekliyor, bol atlamalı, zor çeviri
X002	tezler	13 adet	doc, pdf	very good	www.cmpa.boun.edu.tr	kısa tezler sadece, PDF handling sorunu...

Figure 3.1.1 - Continued

3.1.1.1. The reference links for the texts

Below, we put the reference links and corresponding ID numbers for Figure 3.1.1 for the text files we separated from the BOUN Corpus.

- Arthur Clarke, 2001 A Space Odyssey, B037
 - English, <https://drive.google.com/file/d/1w8tSVWOEDt5pcSQHxQ8WuVbkgurWfvCi/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1Q33CFPTvcmmhsz3TI6vHuaMMwi11x3F/view?usp=sharing>
- Lev Tolstoy, Anna Karenina, B019
 - English, https://drive.google.com/file/d/1p4Cb2WqdCyEmX5pGElfpz-Qv_kk5hQUR/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1Owk_wbgSk5vK-tPpCcO0USUSbyCksGMM/view?usp=sharing
- J.K. Rowling, Harry Potter and the Goblet of Fire, B004
 - English, <https://drive.google.com/file/d/1n7ncn725A4HRtH6wERUz0zGURfX9v0qW/view?usp=sharing>

- Turkish, https://drive.google.com/file/d/1JJN1zJSVSs9vEgS_L4SKr_PxsTu2Yol9/view?usp=sharing
- J.K. Rowling , Harry Potter and the Prisoner of Azkaban, B003
 - English, https://drive.google.com/file/d/1YIhiUFUtrcpD4HvTrqBFeECOLB_LTOMt/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1nWrNOqPZkVMAvkO6SpGcZT3xQrV_HDto/view?usp=sharing
- David Eddings , Enchanters' End Game, B036
 - English, <https://drive.google.com/file/d/1MuZfzvcfLLPzoiNtgYrQP9V-Yl9CXY13/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1JAcI85P8NbMu35MqtxOZHJn7T5oIpimi/view?usp=sharing>
- Bilkent Yönetmelik, T001
 - English, <https://drive.google.com/file/d/1tAH5PXPcxI5xOmIr1SanulhgSH2ZWD28/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1JAcI85P8NbMu35MqtxOZHJn7T5oIpimi/view?usp=sharing>
- Voltaire , Candide, B023
 - English, <https://drive.google.com/file/d/1nrPyq-L9rG7GbeCd-6NzydL6OMpgnTcQ/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/13vHgFY9MfW-H-Qg8cYiaBJ8X0lfTl3sWO/view?usp=sharing>
- Aristoteles , The Athenian Constitution, B020
 - English, <https://drive.google.com/file/d/12829h9FmEDi3RFRgBnPIvi1aUPD0csDF/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1KbQi4OrJAeUQrVCz62wlbH8RMDfxoI5y/view?usp=sharing>
- David Eddings , Queen of Sorcerers, B033
 - English, https://drive.google.com/file/d/1HYHXWdFm0h_OC69nqRi7AGvqKtgBxpju/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1HYHXWdFm0h_OC69nqRi7AGvqKtgBxpju/view?usp=sharing
- David Eddings , Castle of Wisardy, B035
 - English, <https://drive.google.com/file/d/1NiDyct4MRVZdXNUwcD2diAc7xVXTxf9/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1RSP099j2BzWlCxn7IcdlCVUrzmt-sOgB/view?usp=sharing>
- Bernard Shaw , Caesar and Cleopatra, B040
 - English, <https://drive.google.com/file/d/1xI7XePmM641gs-5cWKOdh2SQESAP28OI/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1HjSenG5iP2aACHORpMjX5OGGTHg5kS8/view?usp=sharing>
- Dan Brown , The Da Vinci Code, B012
 - English, <https://drive.google.com/file/d/186aanGkwbcft5ZbcDtRD0GYY3DChpY5H/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1utY1-IKPJxdYqEcI4tk1KGj5YlBIjASB/view?usp=sharing>
- Descartes , Discourse on Method, B013
 - English, <https://drive.google.com/file/d/1r7zfxipeQDHIDZNAKkdASmB4NwK1KJPP/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/11fODE7mRJLG-99srxKx2RaQlEca3gL-X/view?usp=sharing>
- Lenin , State and Revolution, B025
 - English, <https://drive.google.com/file/d/1K5i4UnLMvp3JCmXJEIZDWCZ9E206EtyU/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1ZI-HisCn6X20DjaZPfnjb9liAdzy03a6/view?usp=sharing>
- Dostoyevski , Ecinniler, B055
 - English, <https://drive.google.com/file/d/1gRRb9NqArL2RTMHDWz-N6Kzbr27pn3hq/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1xr_WUbPIPLzz6UqX_A9np5RHWLF_G2HR/view?usp=sharing
- Tolstoy , Master and The Man, B030
 - English, https://drive.google.com/file/d/1B53_LEEwdog4wIqCJYGKdTHdGnUQZf5L/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/19YFQh6u5wUIZr6gojIQZLlFQg1xIILY1/view?usp=sharing>

- Erhan Sigorta, T002
 - English, <https://drive.google.com/file/d/1eosV5SA2FJHGcgScfOv4uol3Js4NCkcH/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1QHUzx0y34SdHXVUoLYMoVq_wa3XSorhq/view?usp=sharing
- Honore de Balzac , Eugenie Grandet, B056
 - English, https://drive.google.com/file/d/1z9QP-5Yj74tehXvCnnRYB8_ZhDdqVLkX/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1B4MW1LfgGvT6cQVIHuIQzZZVQCrCzQed/view?usp=sharing>
- J.K.Rowling , Philosopher's Stone, B001
 - English, https://drive.google.com/file/d/1Q5rjGPMFkLuEjWv2_eY3btQCeutZs8-T/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1QZdBf3P0Rk_sAYLojZdvZn6sfW94LJbY/view?usp=sharing
- Eric Roth , Forrest Gump, X001
 - English, https://drive.google.com/file/d/1PPeiMrurj8s0jNs7VEWJeRnEw3N1-_Vr/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1rjxpg4fYHQYvke0AuhcRqNNYAqWQs1lq/view?usp=sharing>
- Henry David Thoreau , On the Duty of Civil Disobedience, B018
 - English, <https://drive.google.com/file/d/1BsOhGIXLqx19KIwPeDQMqlhcCSzy2By5/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1JMDX0SqzMvQq9ADxpaGNzwQV1GZR5G3y/view?usp=sharing>
- Harun Yahya , Book Collection
 - English, <https://drive.google.com/file/d/1iiDRKN1r5a6Od4iKVCfHXUW42c8jbLji/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1QkIhMbTvbGM1jEVVru4B62pxmFad7wJ/view?usp=sharing>
- Stephen King , Pet Sematary, B011
 - English, https://drive.google.com/file/d/12P_DhuMTqr4PkcP5n0ji2kIRJnFG-g5K/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1yTT_rElNwWW7DN6d-tjyXOXNTCCc7_hh/view?usp=sharing
- Hotel Baykal, T008
 - English, https://drive.google.com/file/d/1qe1VuogATKnvH4R-52Ka_zhRDmj2IZdF/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/143bgUXh-Kj3MN7eEFy1X3wNcOgkbVFtR/view?usp=sharing>
- BOUN , Advanced English, T003
 - English, <https://drive.google.com/file/d/1Fa7fyL88MXCXwnwCWfo2sHPkXmywUrNH/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1XHAKBbX_M3hbiRasXHyJDgDdWTSInbPn/view?usp=sharing
- Moliere , Imaginary Invalid, B052
 - English, <https://drive.google.com/file/d/1L1U8R66SdAIaDcj4lVMliaOKsfyCmDnk/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1eONy7mEqj41MXw7EULAJU-0IaiBPmbAm/view?usp=sharing>
- Goethe , Iphigenia in Tauris, B050
 - English, https://drive.google.com/file/d/1tUVM37oVmyVFe_Z8mXHVUT5NZtH3ESLI/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1EoL8feMrk4K7ZahYRMf5n85a_wHtspjRD/view?usp=sharing
- Mehmet Akif Ersoy , The Turkish National Anthem, T009
 - English, https://drive.google.com/file/d/1pYZmlSiSrGHbaAWnip01_mFP-A1_wBqD/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1unwm1OB8JZqeJa7IlrMFPmG_KnF1tvhPb/vi ew?usp=sharing
- Tolstoy , The Death of Ivan Ilych, B031
 - English, <https://drive.google.com/file/d/1EhUIPDCCpDdlUSx6tHjnR-UeU3vdsZvs/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1GjilCuD1qJzby-Y0SzU9GS18ydhm6qoo/view?usp=sharing>
- Dostoyevski , The Gambler, B049

- English, <https://drive.google.com/file/d/1NLp6enDMg0Fk9Jie8TL-HM9qde54FERF/view?usp=sharing>
- Turkish, https://drive.google.com/file/d/1QMpT0gq8asPfmaPbsohYub-_nWbAHBQ2/view?usp=sharing
- TEB , Jewellers Block Insurance,
 - English, https://drive.google.com/file/d/12owzJnaTmEbcVVPm1up4XT3Zj_C7YfRa/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/13o14qvkvACYjzvY69mc6q988e0Slf0CV/view?usp=sharing>
- Shakespeare , Macbeth, B010
 - English, https://drive.google.com/file/d/1nMYHxpXgWR_o6AoVk3OjWwS3DM7jl3Hw/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1TRm-Ep40XoU0cG1QE0YvDiqx6-JaR2CD/view?usp=sharing>
- David Eddings , Magician’s Gambit, B034
 - English, <https://drive.google.com/file/d/10CGk37ql1PpboauMfnbKXdVceImqxfz/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1dHfgvWzh8MUxarjTGSkk9wnfQWgtM-y/view?usp=sharing>
- The Wachowskis , Matrix, X001
 - English, <https://drive.google.com/file/d/1pb3Vshf6fqwoa2Og09iRMgmUpe5Tefo/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/15TkRSqjt8GShPHm2JccQZejW0P2NZarD/view?usp=sharing>
- Oscar Wilde , The Happy Prince, Tales, B060
 - English, https://drive.google.com/file/d/1-Afj3BkJbS9yELhWY6-cn0OZGx_hio_x/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1LwRM9lkXQgtGWiPO6YzkwQ8X0YeRGK6y/view?usp=sharing>
- Musteri Iliskileri
 - English, <https://drive.google.com/file/d/1cX27D0PWG6tRWUnalL7yV-TEyGEBQJYM/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1e07zs5_0UElvPh5px0qECXvnv6FAiX-k/view?usp=sharing
- Bob Peterson , Finding Nemo, X001
 - English, https://drive.google.com/file/d/1n2COa0WzkuBE_S4NQflwxwJ94Kj_Cno_/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/16EDImzkSgYqf92YzuFiJ4ICLRuVnVZK_/view?usp=sharing
- Lucretius , On The Nature Of Things, B029
 - English, https://drive.google.com/file/d/1_2scOp53LMln-vXq3OKF3p9fk6x_bDyM/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1dYBtayFCN5dQ5O2Hz7niOPWMSYLM5mI6/view?usp=sharing>
- Anatole France , Penguin Island, B058
 - English, <https://drive.google.com/file/d/12Gvr7NhGTUNeqClAeQCVYQOsg3TnIRoN/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1LYvsUocNMuqzwx3p6LW1OcWL28hqnd/view?usp=sharing>
- Anna Sewell , Pollyanna, B044
 - English, https://drive.google.com/file/d/1iKG3_x1Ou_TBUGLiA2IoQupwmRBbBCr3/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1hPL8CZ571AiJJ2bF9BhzUOaxFNMsiPSM/view?usp=sharing>
- Arthur C. Clarke , Rama, B038
 - English, https://drive.google.com/file/d/1z9E97GkPXVK9fdu_Bq9AATWf1frXPjfv/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1SCO5C8ay7NJvC5SJzl2Yky22BGAHkvOd/view?usp=sharing>
- Arthur C. Clarke , Rendezvous With Rama, B039
 - English, https://drive.google.com/file/d/1akmbbg9F6__LTE7_OOc_Ibz1LLWHbOTj/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1nR0xunbW5_EqBhoRfBwmyMIznIjLuB4e/view?usp=sharing
- Turgenev , Rudin, B047
 - English, <https://drive.google.com/file/d/1qhXGCzDp8RXzMOx4RmyWOCJZ2w5rCVz1/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1vOeFc08vIbgurKi71hym4fMi9R8nrOWR/view?usp=sharing>

- Gaston Leroux , Mystery of Yellow Room, B053
 - English, https://drive.google.com/file/d/1dHXmel-PEjjySZjko2rG6XnQFv-jJb_5/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1pibQgTQuvXhRj4zMaNrYBW7qkEd2m_lH/view?usp=sharing
- Lokman Cayci , Savas Danslari, T013
 - English, https://drive.google.com/file/d/1u9RQ_gFqMehYjxSV-hDwT9rocGODdW26/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/1Q2FcXh6rHjIVpW-3eZiUujt1lArSXQhv/view?usp=sharing>
- Halil Dagdelen , Sen Suclu Degilsin, T014
 - English, https://drive.google.com/file/d/1PxdnFNsiB1k6bd4ihDLdHiEgoYuTGc_x/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/15gxqz7iNZLdouVz4lzZ1oWk2_36248u9/view?usp=sharing
- J.K.Rowling , Chamber of Secrets, B002
 - English, <https://drive.google.com/file/d/12LCzaOsAjN1TMdtSB7ZvjZu5P0X3b6O/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1EzppI_gO9cBG1oGO1NaB2mwrwq6sq_fr/view?usp=sharing
- Plato , Socrates' Defence, B026
 - English, <https://drive.google.com/file/d/1qlhghqxiHILA5GBSAqAYAvbUmYTVTO4J/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/17YnoS5tX8FBnp1wOltAjBd7sI6uyqFoi/view?usp=sharing>
- Peter Schlemihl , Chamisso, B059
 - English, <https://drive.google.com/file/d/1pcmiFR4IihiBABcdjIRXS4rp359we-uR/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1_qkg_ptmii6QsyiR6pxjdvvB2_GDbKqa/view?usp=sharing
- Platon , Statesman, B015
 - English, <https://drive.google.com/file/d/1IaqQMlxB3b6j0tPlQ51plDw-KSHvM3LD/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1ZzYdlSLD06otDdcnRmydr_z93w4-9B91/view?usp=sharing
- Tolstoy , Crime and Punishment, B061
 - English, https://drive.google.com/file/d/1LMXymWoHX_QTuQtHcJA4cM_wHWKwc8NI/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1Hs-2m3k8r2fydcR0pBS_K2KgtkMMLBXd/view?usp=sharing
- The Peach, T011
 - English, <https://drive.google.com/file/d/1AsxPX6Mpz-qfWZT4QgRMzmtym1c2vAS4/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1FsUEOvLy9XfyHIZuvTdrUnlOYs_6gvD3/view?usp=sharing
- Thais , Anatole France, B045
 - English, <https://drive.google.com/file/d/1dOth2FFkJHlZjXH3tVCYR4UwLOddsbGc/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1Im1mxGUwFkCQnSDNb5Vnq0JnxHLUnA8m/view?usp=sharing>
- Harun Yahya , The Golden Age, B071
 - English, https://drive.google.com/file/d/1BI5roylpZr509D0fNQ82qqMAlrG_ARy7/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1IF0QlNv8BhIINOVBIlqGAwVP3zU-_7tC/view?usp=sharing
- Robert Louis Stevenson , The Merry Men,
 - English, <https://drive.google.com/file/d/1W6RCAjFfSpKTjhJVLQB-R5P8WdKB2NDJ/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/17sMffMH9GYskyk_UR3yGQRWnT1jD8MbT/view?usp=sharing
- Platon , Republic, B021
 - English, https://drive.google.com/file/d/1zbwxwCi23Lg86-McZDmxaemmOdW_JxUU/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1c-mIMMmluTkcih1CpaSKCwXaQf-3N_kG/view?usp=sharing
- James Cameron , Titanic, X001
 - English, <https://drive.google.com/file/d/1-3116-GhJmMTbnfFpC-ao7I2YGTGe8z0/view?usp=sharing>

- Turkish, https://drive.google.com/file/d/1uJhASadtbWbp3BNvavOvuJu_z8nXdoSl/view?usp=sharing
- Mark Twain , The Advantures of Tom Sawyer, B022
 - English, https://drive.google.com/file/d/1dENV8FmtCqWggempCc2fmc3A7i_6Rl1d/view?usp=sharing
 - Turkish, <https://drive.google.com/file/d/115aecpf20CF8MMVqjHlQI1g4NtBNuJpO/view?usp=sharing>
- Guillermo Arriaga , 21 Grams Script, X001
 - English, <https://drive.google.com/file/d/1YazcVojRyDOvb6FFBSapdmWT09BwtsUm/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1CQAxlxIzl0QqcJ7YXwvW0MjbNuubxqJu/view?usp=sharing>
- Jack London , The Call of the Wild, B054
 - English, <https://drive.google.com/file/d/1v6-Q5tVDWMUyCuf1ttillPRnzhMBI8Yl/view?usp=sharing>
 - Turkish, <https://drive.google.com/file/d/1IOUjCsxkVRzgWU0HbgwChG7NtqPvQlLL/view?usp=sharing>
- Plesk
 - English, https://drive.google.com/file/d/1cl_w5jOceHunsGAhWuHWifr-HRVB-E77/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1mxxI0HyGpI1EWH6mgG8ZoQD3Ai_kht3t/view?usp=sharing
- Marcus Tullius Cicero , Friendship and Old Age, B027
 - English, <https://drive.google.com/file/d/1FotL74NwpUhGhbLiUjPhzQH-YGTTSJfT/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1FTPb3n7ZhVJkJf6DKKmOtiwxNsW5X_t2/view?usp=sharing
- Yazilikaya, T012
 - English, <https://drive.google.com/file/d/1z64pNv6PvUgmosOMes212-DCCumpR8WA/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1-tV1XHKKHJfycnkk6xn9Xazsat0WM_3x/view?usp=sharing
- Francis Bacon , New Atlantis, B014
 - English, <https://drive.google.com/file/d/1Aol2gFnOwBcPFFMwULDYciTO9mkV0rOf/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1ScRJN_xDHDp574EGazzSAPrLeasiyqR7/view?usp=sharing
- Dostoyevski , Notes From Underground, B017
 - English, <https://drive.google.com/file/d/1l3S1pnYaDDE3IOB7hHFX2lE8Lh2pAAia/view?usp=sharing>
 - Turkish, https://drive.google.com/file/d/1gPtIKqwShzNuxnhCxUasK4OX7b32_ukR/view?usp=sharing
- Stephen King , The Green Mile, B028
 - English, https://drive.google.com/file/d/1zGu1Hmq-KpQ0PR_-dMPHiGzc2FfgXsCE/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1n5O5twgM9031hKOfkBma_fLbIJizQXrW/view?usp=sharing
- J.K.Rowling , The Order of the Phoenix, B005
 - English, https://drive.google.com/file/d/159VXtW-p6ieIhJ7mJVW_nnxfehMA2cDL/view?usp=sharing
 - Turkish, https://drive.google.com/file/d/1R8aD8Uxxx_3P60Et_pP3FdEV4SYcJJEQ/view?usp=sharing

3.2. Installation of Moses and Giza

Moses [2] is a statistical machine translation tool that is widely used in this area. We followed the instructions on Moses’s website [10] when we installed it on our computers.

Moses runs on a Linux based system. Because we both used Windows, we needed to

install Ubuntu subsystem on Windows. For installing a subsystem, we need to enable it on our computers by running the command below on Windows PowerShell.

```
$Enable-WindowsOptionalFeature -Online -FeatureName  
Microsoft-Windows-Subsystem-Linux
```

After enabling it, we downloaded Ubuntu18.04 to this subsystem and installed it with the commands below.

```
$Invoke-WebRequest -Uri https://aka.ms/wsl-ubuntu-1804 -OutFile  
Ubuntu.appx -UseBasicParsing  
$Add-AppxPackage .\Ubuntu.appx
```

After successfully installing Ubuntu subsystem, we needed to install the dependency libraries of Moses. This is the command we used to install the necessary libraries.

```
$sudo apt-get install [package name]
```

The libraries are:

```
g++  
git  
subversion  
automake  
libtool  
zlib1g-dev  
libc++-dev  
libboost-all-dev  
libbz2-dev  
liblzma-dev  
python-dev  
graphviz  
imagemagick  
make
```

```

cmake
libgoogle-perftools-dev (for tcmalloc)
autoconf
doxygen

```

After installing the packages, we needed to install bjam as Moses requires bjam for running. To upload bjam the commands are:

```

$wget https://dl.bintray.com/boostorg/release/1.64.0/source/boost_1_64
_0.tar.gz
$tar zxvf boost_1_64_0.tar.gz
$cd boost_1_64_0/
$./bootstrap.sh
$./b2 -j4 --prefix=$PWD --libdir=$PWD/lib64 --layout=system
link=static install || echo FAILURE
$./bjam --with-boost=~/.workspace/temp/boost_1_64_0 -j4

```

After executing the commands above, we started to have problems on running b2. In the installation guide the last command is recommended to run bjam because Moses needs to know where is the boost libraries are installed. But despite trying many versions of the library location, we couldn't manage to run the system.

Because there aren't enough resources about the problems about Moses on the Internet we couldn't find a solution there. Also, the installation guide is the only guide on Internet about the installation of Moses.

To solve our problem we consulted to Buse Buz, a former student of Tunga Gungor who worked with Moses before. With her help we were able to install Moses at last. The solutions we tried are:

1. Downloading and extracting the tar file again for boost installation.
2. Running the ./bjam command on the same directory of the boost library file because we couldn't give the address as argument.
3. Cloning the Moses file from Moses GitHub repository [11]. The repo also contains

the boost library for the bjam command.

After applying the solutions above, we were able to run the bjam command that installs the Moses. The third solution is the most essential one as the installation guide doesn't mention that we needed to download Moses at some point from an external resource. We assumed Moses will appear after we did the steps on the installation guide [10]. To run the Moses run the bjam command below.

```
$. /bjam -j4
```

We didn't need other bjam commands in the installation guide. The command takes a long time to run. After running this command a mosesdecoder directory needs to appear otherwise it means there is an error.

To run Moses [2], a word alignment tool is necessary. We used Giza++ as it works with Moses system and it was recommended by our professor. To install Giza++, we followed the steps in [12]. The commands are listed below.

```
$git clone https://github.com/moses-smt/giza-pp.git
$cd giza-pp
$make
$cd ~/mosesdecoder
$mkdir tools
$cp ~/giza-pp/GIZA++-v2/GIZA++ ~/giza-pp/GIZA++-v2/snt2cooc.out
~/giza-pp/mkcls-v2/mkcls tools
```

After successfully executing these commands, Moses is ready for running.

3.3. Running Moses

To run the moses we follow the instructions at its website [12]. Since moses does not have a Turkish option, we make the tokenization and truncating as English for both languages.

3.3.1. Corpus Preparation

Corpus preparation was the first step of our project. The corpus files' names are "BU_en.txt" and "BU_tr.txt"

Some following files created as txt type but it was a rookie mistake and causes no error.

For tokenization we used the following commands:

```
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < ~/Bu_en.txt > ~/token.en.txt
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < ~/Bu_tr.txt > ~/token.tr.txt
```

For truecasing we first trained truecaser for some statistics with the following commands.

```
~/mosesdecoder/scripts/recaser/train-truecaser.perl -model ~/train_truecaser.en.txt
-corporus ~/token.en.txt
~/mosesdecoder/scripts/recaser/train-truecaser.perl -model ~/train_truecaser.tr.txt
-corporus ~/token.tr.txt
```

For truecasing we used the following commands.

```
~/mosesdecoder/scripts/recaser/truecase.perl -model ~/train_truecaser.en.txt <
~/token.en.txt > ~/true.en
~/mosesdecoder/scripts/recaser/truecase.perl -model ~/train_truecaser.tr.txt <
~/token.tr.txt > ~/true.tr
```

For cleaning we used the following commands.

```
~/mosesdecoder/scripts/training/clean-corpus-n.perl ~/corpus/true tr en ~/clean 1
```

3.3.2. Language Model Training

For creating language model and binarise it we used the following commands.

```
mkdir ~/lm
cd ~/lm
~/mosesdecoder/bin/lmplz -o 3 <~/true.en > arpa.en.txt
~/mosesdecoder/bin/build_binary arpa.en.txt blm.en
```

This step can be checked with the following command, it gives a result trying to convert the sentence written.

```
echo "is this an English sentence ?" | ~/mosesdecoder/bin/query blm.en
```

3.3.3. Training the Translation System

Training takes a lot of time. It took more than 6 hours, may be fasten with -cores argument.

```
mkdir ~/working
cd ~/working
nohup nice ~/mosesdecoder/scripts/training/train-model.perl -root-dir train -corpus
~/clean -f tr -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm
0:3:$HOME/lm/blm.en:8 -external-bin-dir ~/mosesdecoder/tools > training.out
```

3.3.4. Tuning

Tuning also took about 4 hours. Running it multi-threaded said to be a lot faster but related commands did not work for our computers.

```
cd ~/corpus
wget http://www.statmt.org/wmt12/dev.tgz
tar zxvf dev.tgz
```

We used news-test2008 for tuning, so we tokenised and truecased it first.

```

cd ~/
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < dev/news-test2008.en > news-
test2008.tok.en
~/mosesdecoder/scripts/recaser/truecase.perl -model train_truecaser.en.txt < news-
test2008.tok.en > news-test2008.true.en
~/mosesdecoder/scripts/recaser/truecase.perl -model train_truecaser.tr.txt < news-
test2008.tok.en > news-test2008.true.tr

```

Then we launched the tuning process.

```

cd ~/working
nohup nice ~/mosesdecoder/scripts/training/mert-moses.pl /news-test2008.true.tr
~/news-test2008.true.en ~/mosesdecoder/bin/moses train/model/moses.ini -mertdir
~/mosesdecoder/bin/ > mert.out &

```

3.3.5. Testing

Moses can be run with the following command then typing a sentence to try.

```
~/mosesdecoder/bin/moses -f ~/working/mert-work/moses.ini
```

In order to make it start quickly, we binarised the phrase-table and lexicalised reordering models with the following commands.

```

mkdir ~/working/binarised-model
cd ~/working
~/mosesdecoder/bin/processPhraseTableMin -in train/model/phrase-table.gz -nscores
4 -out binarised-model/phrase-table
~/mosesdecoder/bin/processLexicalTableMin -in train/model/reordering-table.wbe-msd-
bidirectional-fe.gz -out binarised-model/reordering-table

```

To measure this, we used another parallel data set (the test set). We used harry potter goblet of fire

```
cd ~/corpus
```

```

~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < ates_kadehi_en.txt > ates.tok.en
~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l fr < ates_kadehi_tr.txt > ates.tok.tr
~/mosesdecoder/scripts/recaser/truecase.perl -model train_truecaser.en.txt < ates.tok.en
> ates.true.en
~/mosesdecoder/scripts/recaser/truecase.perl -model train_truecaser.tr.txt < ates.tok.tr
> ates.true.tr

```

The model that we've trained can then be filtered for this test set, meaning that we only retain the entries needed to translate the test set. This will make the translation a lot faster.

```

cd ~/working ~/mosesdecoder/scripts/training/filter-model-given-input.pl filtered-
ates mert-work/moses.ini ~/ates.true.tr -Binarizer ~/mosesdecoder/bin/processPhraseTableMin

```

We tested the decoder by first translating the test set (takes a wee while) then running the BLEU script on it:

```

nohup nice ~/mosesdecoder/bin/moses -f ~/working/filtered-ates/moses.ini <
~/ates.true.tr > ~/working/ates.translated.en 2> ~/working/ates.out
~/mosesdecoder/scripts/generic/multi-bleu.perl -lc ~/ates.true.en <
~/working/ates.translated.en

```

This gives us a BLEU score of 17.02

3.4. Data Preprocessing

When we tested our trained system with a part of the training data, we get 17.02 BLEU score. Since this is not a reliable result, we divided the corpus into three parts as 550K, 68800, 69470 lines long, for train, tune and test respectively. With 3-gram model and without any preprocessing it gives BLEU score of 0.13 which is noticeably low, which leads us to data preprocessing to get better results.

3.4.1. Stemming words

For English stemming we used PorterStemmer from Python's NLTK library [13], since it is commonly used and reliable with results. Algorithm behind the PorterStemmer is referenced as [14].

For Turkish Stemming we used Zemberek library [15], which is also common to use in Turkish NLP works. We followed this medium tutorial [16] to use Zemberek.

Stemming improve the score to 0.35 from 0.13, so we used stemming in our next steps.

Turkish Stemming

```
from typing import List
from jpyype import JClass, JString, getDefaultJVMPath, shutdownJVM, startJVM,
    java
import string

def file_stemmer_turkish(read_filename, write_filename, morphology):
    file_read = open(read_filename, encoding="utf8")
    file_write = open(write_filename, "w+", encoding="utf8")
    lines = file_read.readlines()
    for words in lines:
        words = words.translate(str.maketrans('', '', string.punctuation))
        analysis: java.util.ArrayList =
            (morphology.analyzeAndDisambiguate(words).bestAnalysis())
        pos: List[str] = []
        for i, analysis in enumerate(analysis, start=1):
            f'\nAnalysis {i}: {analysis}',
            f'\nPrimary POS {i}: {analysis.getPos()}',
            f'\nPrimary POS (Short Form) {i}: {analysis.getPos().shortForm}'

        pos.append(
```

```

        f'{str(analysis.getLemmas()[0])}'
    )
    result_file.write(f'{" ".join(pos)}\n')

ZEMBEREK_PATH = r'zemberek-full.jar'
startJVM(getDefaultJVMPath(), '-ea', '-Djava.class.path=%s' %
        (ZEMBEREK_PATH))
TurkishMorphology = JClass('zemberek.morphology.TurkishMorphology')
morphology = TurkishMorphology.createWithDefaults()
file_stemmer_turkish("read_filename.txt", "write_filename.txt", morphology)

```

English Stemming

```

import string
import nltk
from nltk.stem.porter import PorterStemmer
porter_stemmer = PorterStemmer()

def file_stemmer(read_filename, write_filename):
    file = open(read_filename, "r", encoding="utf8")
    file_write = open(write_filename, "w+", encoding="utf8")
    for line in file.readlines():
        # First Word tokenization
        nltk_tokens = nltk.word_tokenize(line)
        # Next find the roots of the word
        for w in nltk_tokens:
            file_write.write(porter_stemmer.stem(w) + " ")
        file_write.write("\n")

```

3.4.2. Removing Punctuation

We removed punctuation from stemmed data.

Removing punctuation decreased the score to 0.26 from 0.35 so we did not use it for the next steps.

Removing Punctuation

```
def file_wout_punc(read_filename,write_filename):
    file = open(read_filename,"r", encoding="utf8")
    file_write = open(write_filename,"w+", encoding="utf8")
    file_lines = [''.join(c for c in s if c not in string.punctuation) for s
                  in file.readlines()]
    file_write.write(file_lines + "\n")
```

3.4.3. Stopword Removal

We used NLTK's stopwords for removing stopwords. NLTK Corpus has stopword list for 16 different languages which includes English and Turkish. After stopword removal, we also stemmed the data.

Stopword removal is also decreased the score to 0.13 from 0.35, hence used only stemming in these steps.

Stopword Removal

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')

def remove_stopwords(read_filename,write_filename,lang):
    file = open(read_filename,"r", encoding="utf8")
    file_write = open(write_filename,"w+", encoding="utf8")
    for line in file.readlines():
        # First Word tokenization
```

```

nltk_tokens = nltk.word_tokenize(line)
#Next find the roots of the word
words = [word for word in nltk_tokens if not word in
          stopwords.words(lang)]
sentence = (" ").join(words)
file_write.write(sentence)
file_write.write("\n")

```

3.4.4. Shuffling Data

Our corpus contains many books, texts and data, but they are concatenated. Splitting such data with an order is not a good way to train the system since system is only trained by some books, and testing with another part of corpus may lower the results since system may not recognize the words. So we shuffled the data in parallel manner, then stemmed and retry the system.

Shuffling increased the score to 0.41 from 0.35.

Parallel Shuffling

```

import numpy as np
import random
import math
N = 100000

def file_shuffle_parallel(en_file,tr_file,en_file_shuffled,tr_file_shuffled):
    en_file = open(en_file,"r", encoding="utf8")
    tr_file = open(tr_file,"r", encoding="utf8")
    tr_file_shuffled = open(tr_file_shuffled,"w+", encoding="utf8")
    en_file_shuffled = open(en_file_shuffled,"w+", encoding="utf8")
    en_lines = en_file.readlines()
    tr_lines = tr_file.readlines()
    count = len(en_lines)

```

```

indices = np.array(range(N))
random.shuffle(indices)
for i in indices:
    if i == N:
        a = np.column_stack(np.array([en_lines[i*math.ceil(count/N):],
                                       tr_lines[i*math.ceil(count/N):]]))

        np.random.shuffle(a )
        en_file_shuffled.writelines(a.T[0])
        tr_file_shuffled.writelines(a.T[1])
    else:
        a = np.column_stack(np.array([en_lines[i*math.ceil(count/N):(i+1)
                                       * math.ceil(count/N)], tr_lines[i*math.ceil(count/N):(i+1) *
                                       math.ceil(count/N)]]))
        np.random.shuffle(a)
        en_file_shuffled.writelines(a.T[0])
        tr_file_shuffled.writelines(a.T[1])

```

3.4.5. Changing Train Data

We decreased our training data size to 480K lines from 550K lines, in order to see the difference.

Decreasing train data size, decreased the score to 0.39 from 0.41 as expected.

3.4.6. 4-gram Language Model

So far, we used 3-gram language models in our system. Expecting an increase in results, we tried 4-gram language model with our stemmed data.

This increased the score to 0.45 from 0.41.

3.5. NMT

After experimenting on Moses, we decided to work also with NMT, we read and analyze Nemasus [17] and Marian [18]. Since Marian has more comprehensive tutorial we decided to work with Marian on NMT.

Unfortunately, our computers cannot handle any NMT tool in terms of CPU and RAM, so we did not work with NMT tools, instead only worked with SMT.

4. RESULTS

4.1. Turkish-English Parallel Corpus

We have examined the previous works on parallel corpora. We have seen that although there are multiple aligned corpora on different categories, there isn't any organized, extensive corpus yet. Also, we have searched for the algorithms that we can use for sentence alignment. After our research, we decided to improve [1] which we already have to be a more categorized corpus.

The Boun [1] corpus is divided into smaller sections that contain separate books. You can see the resulting books and their line counts from tables 4.1, 4.2 and 4.3. Figure 4.1 shows the data type - line number relation for some part of the corpus that we separated and processed. Figure 4.2 shows the distribution of number of text types over genres for example we separated 49 books from the corpus.

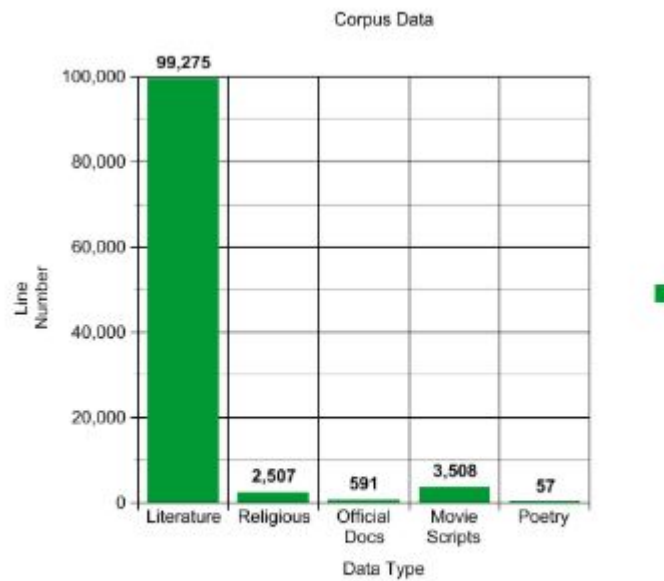


Figure 4.1. Data type - Line number graph

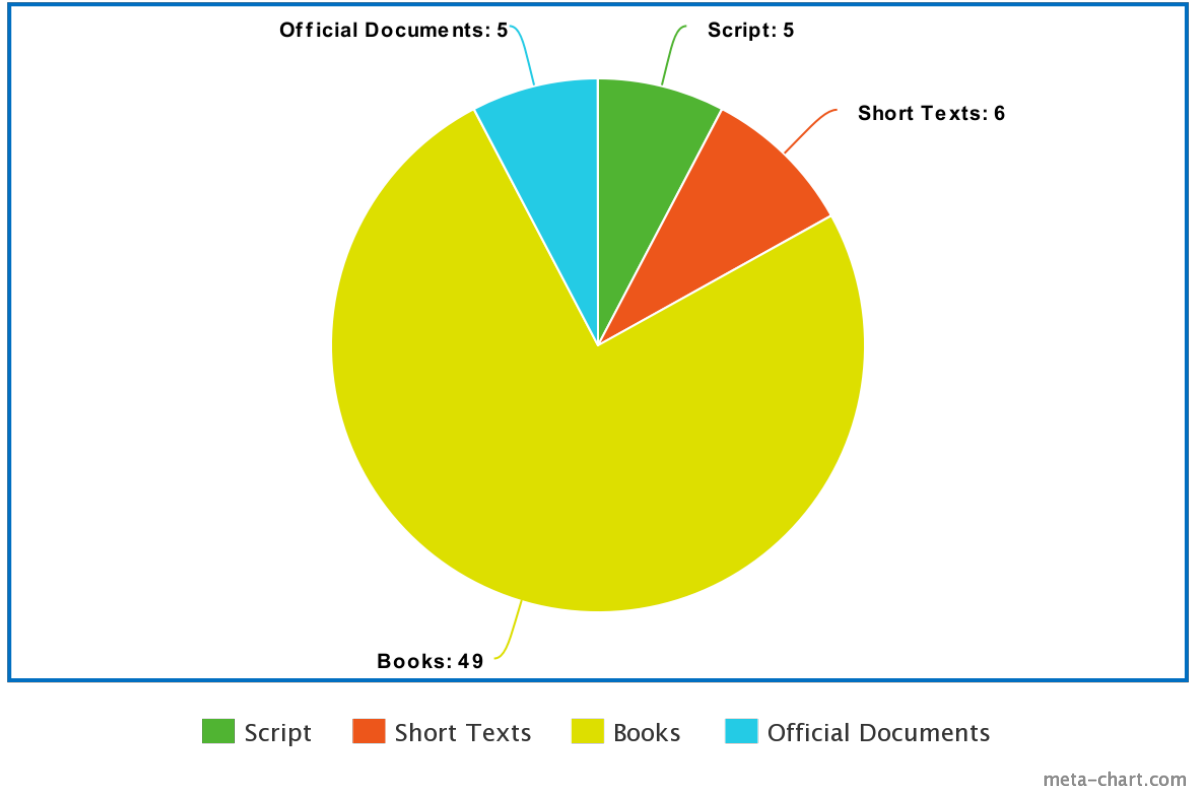


Figure 4.2. Distribution over genres

Table 4.1: Information about the partitioned books

Title	Author	Line Count
2001 A Space Odyssey	Arthur Clarke	1194
Anna Karenina	Lev Tolstoy	6713
Goblet of Fire	J.K. Rowling	10051
Prisoner Of Azcaban	J.K. Rowling	2922
Enchanters' End Game	David Eddings	6626
Bikent Yönetmeliği	Bilkent Üniversitesi	81
Candide	Voltaire	795
The Athenian Constitution	Aristoteles	658
Queen of Sorcerers	David Eddings	6965
Castle of Wisardy	David Eddings	2915
The Da Vinci Code	Dan Brown	4280
Discourse on Method	Descartes	194

Table 4.1 Continued

State and Revolution	Lenin	1021
Ecinniler	Dostoyevski	2830
Master and Man	Tolstoy	421
Erhan Sigorta	Erhan Sigorta	104
Eugenie Grandet	Honore de Balzac	2010
Philosopher's Stone	J.K.Rowling	3398
Forrest Gump	Eric Roth	975
On the Duty of Civil Disobedience	Henry David Thoreau	2345
Book Collection	Harun Yahya	1937
Pet Sematary	Stephen King	2760
Hotel Baykal	Hotel Baykal	9
Adcanved English	BOUN	95
Imaginary Invalid	Moliere	882
Iphigenia in Tauris	Goethe	222
The Turkish National Anthem	Mehmet Akif Ersoy	8
The Death of Ivan Ilych	Tolstoy	915
The Gambler	Dostoyevski	1305
Jewellers Block Insurance	TEB	106
Macbeth	Shakespeare	782
Magician's Gambit	David Eddings	3652
Matrix Script	The Wachowskis	807
The Happy Prince,Tales	Oscar Wilde	261
Musteri İlişkileri	-	172
Finding Nemo	Bob Peterson	965
On The Nature Of Things	Lucretius	461
Penguin Island	Anatole France	2387
Pollyanna	Anna Sewell	593
Rama	Arthur C. Clarke	2917
Rendezvouz With Rama	Arthur C. Clarke	2531

Table 4.1 Continued

Rudin	Turgenyev	1845
Mystery of Yellow Room	Gaston Leroux	2368
Savas Danslari	Lokman Cayci	34
Sen Suclu Degilsin	Halil Dagdelen	15
Chamber of Secrets	J.K.Rowling	3547
Socrates' Defence	Platon	333
Chamisso	Peter Schlemihl	285
Statesman	Platon	429
Crime and Punishment	Tolstoy	2399
The Peach	-	18
Anatole France	Thais	1770
The Golden Age	Harun Yahya	570
The Merry Men	Robert Louis Stevenson	247
Republic	Platon	776
Titanic Scenario	James Cameron	1199
The Adventures of Tom Sawyer	Mark Twain	1075
21 Grams Script	Guillermo Arriaga	527
The Call of the Wild	Jack London	1327
Plesk	-	26
Friendship and Old Age	Marcus Tullius Cicero	218
Yazilikaya	-	7
New Atlantis	Francis Bacon	421
Notes from Underground	Dostoyevski	1382
The Green Mile	Stephen King	2839
The Order of the Phoenix	J.K.Rowling	3008

4.2. Statistical Machine Translation System

Firstly, we tried the system with a small file to test the commands we found, understand the process and get quicker results. Then we worked on the main corpus where it gives us more accurate results but takes a long time.

We got a BLEU score 17.02 with the main corpus [1], details are on Table 4.2. This is a reasonable result for the first try, since the best result is 30.5 mentioned at [12]. We decide on trying to improve this result then move on to the neural translation systems.

You can see the example translations from table 4.3.

Language Model	Line Count	Tuning	Test Data	BLEU
Boun Corpus	above 700 k	Boun Corpus	Goblet of Fire	17.02

Table 4.2. BLEU Results

Given	Translation
arı mı ?	bees ?
iyi bir çeviri	good a translation
eldiven	gloves
bardak	glass
araştırma	research
adın ne	adın what
isim ne	name what
su	water
içmek	drink
su içmek	water or
kirli yastık	dirty pillow
irem gülüyor	irem laughing
tunga hoca en iyi hoca	tunga teacher best teacher
7 saniye sonra	7 seconds after
7 saniye önce	7 seconds before

Table 4.3. Translation examples

4.2.1. Corpus With Different Sizes And Pre-Processing Techniques

We tried our machine translation system with different sizes of corpus as training data. We also processed the training data in multiple ways as we mentioned in Section 3. From this trials, we got different results. In Table 4.4, there is corresponding ID numbers for corresponding different trials.

Train Size	Tune Size	Test Size	Preprocessing Type	Trial ID
550K Lines	69K Lines	69K Lines	No preprocessing	1
550K Lines	69K Lines	69K Lines	Stem the words in corpus	2
550K Lines	69K Lines	69K Lines	Stem words and remove punctuation	3
550K Lines	69K Lines	69K Lines	Stem words and remove stop-words	4
550K Lines	49K Lines	49K Lines	Stem words and shuffle corpus	5
480K Lines	49K Lines	49K Lines	Stem words and shuffle corpus	6
550K Lines	49K Lines	49K Lines	Stem words and shuffle corpus with 4-gram LM	7

Table 4.4. IDs for different trials of translation system

4.2.2. BLEU Scores Of Different Trials

BLEU (bilingual evaluation understudy) is an algorithm that evaluates the machine translation system's success between two different natural languages. We used BLEU [19] scores evaluating the results of our machine translation system. In Table 4.5, corresponding BLEU scores of our different trials. As you can see, the BLEU scores that we got can be considered very low compared to the usual expected scores. We think the reason for this is that our 3-gram and 4-gram precision scores are very low. Because BLEU score is the geometrical mean of these 4 precision, our global score comes out as low.

Trial ID	BLEU Score
1	BLEU = 0.13, 10.2/0.4/0.0/0.0
2	BLEU = 0.35, 21.9/1.5/0.2/0.0
3	BLEU = 0.26, 12.9/0.9/0.1/0.0
4	BLEU = 0.13, 13.8/0.6/0.0/0.0
5	BLEU = 0.41, 15.6/0.8/0.2/0.1
6	BLEU = 0.39, 18.3/0.7/0.1/0.1
7	BLEU = 0.45, 19.4/0.9/0.2/0.1

Table 4.5. BLEU Scores of different trials

4.2.3. Example Translations With Non-shuffled Corpus

We did our first experiments with the non-shuffled corpus. To create training, tuning and testing data we divided the corpus. In Tables 4.6 and 4.8 contains example input sentences after preprocessing, and the translated results are placed in Table 4.7 and 4.9, respectively. If you look at the sentences, they may not look understandable most because the input data for the system was pre-processed.

Turkish sentence 1: İnsanlar düpedüz dikkatsizlik ediyorlar, sokaklara fırlamışlar güpegündüz, sırtlarında Muggle giysileri bile yok, boyuna dedikodu ediyorlar."

English sentence 1: People are being downright careless, out on the streets in

broad daylight, not even dressed in Muggle clothes, swapping rumors."

Trial ID	Processed Input Sentence-TR
1	İnsanlar düpedüz dikkatsizlik ediyorlar, sokaklara fırlamışlar güpegündüz, sırtlarında Muggle giysileri bile yok, boyuna dedikodu ediyorlar."
2	insan düpedüz dikkat et , sokak fırla güpegündüz , sırt UNK giysi bile yok , boy dedikodu et . "
3	insan düpedüz dikkat et sokak fırla güpegündüz sırt UNK giysi bile yok boy dedikodu et
4	insan düpedüz dikkat et sokak fırla güpegündüz sırt UNK giysi bile yok boy dedikodu et

Table 4.6. Processed Turkish input sentence 1

Trial ID	Translated Output Sentence-EN
1	İnsanlar düpedüz dikkatsizlik ediyorlar get , they criticize them behind sokaklara fırlamışlar güpegündüz get , they criticize them behind sırtlarında Muggle giysileri say " Ugh ! " to them new creation ? " ... contrary , it causes them anxiety . boyuna they ediyorlar and devoting himself to that god . mood " in their paintings , sculptures
2	for düpedüz these of at fırla güpegündüz they sırt , giysi say , then the boy is that what they
3	the düpedüz refer to these in a fırla güpegündüz sırt hi giysi around the same boy they are
4	the life düpedüz dikkat school fırla sırt güpegündüz “ giysi form corrupt televis person boy .

Table 4.7. Translated English output sentence 1

Turkish sentence 2: Dudley’nin ağzı dehşetle açıldı, ama Harry’nin yüreği hopladı.

English sentence 2: Dudley’s mouth fell open in horror, but Harry’s heart gave a leap.

Trial ID	Processed Input Sentence-TR
1	Dudley’nin ağzı dehşetle açıldı, ama Harry’nin yüreği hopladı.
2	UNK ağız dehşet aç , ama UNK yürek hopla .
3	UNK ağız dehşet aç ama UNK yürek hopla
4	UNK ağız dehşet aç UNK yürek hopla

Table 4.8. Processed Turkish input sentence 2

Trial ID	Translated Output Sentence-EN
1	Dudley 'nin ağzı dehşetle açıldı get , they criticize them behind their Harry 'nin yüreği hopladı not know they had said anything unkind
2	, ağız they open , yürek dehşet aç hopla .
3	hi ağız dehşet but to hi yürek hopla
4	except ağız dehşet “ yürek hopla account

Table 4.9. Translated English output sentence 2

4.2.4. Example Translations With Shuffled Corpus

After trying the corpus without shuffling its lines, we decided to try the translation system on the corpus that we shuffled its lines. We expected that the shuffled corpus would give better results because the genres of the books were different and the training data would have all the genres if we shuffled it. In Tables 4.10 and 4.12 contains example input sentences after preprocessing, and the translated results are placed in Table 4.11 and 4.13, respectively.

Turkish sentence 3: Gözleri loş ışığa alıştığında, Floyd geceyi yaşayan yerin tamamen karanlık olmadığını farketti.

English sentence 3: Then Floyd saw, as his eyes grew more accustomed to the fainter illumination, that the night land was not wholly dark.

Trial ID	Processed Input Sentence-TR
5	göz loş ışık alış , floyd gece yaşa yer tamamen karanlık ol farket .
6	göz loş ışık alış , floyd gece yaşa yer tamamen karanlık ol farket .
7	göz loş ışık alış , floyd gece yaşa yer tamamen karanlık ol farket .

Table 4.10. Processed Turkish input sentence 3

Trial ID	Translated Output Sentence-EN
5	ask semidark of alış cast floyd to feel shake server littl away farket .
6	ask semidark of alış of floyd to feel shake server littl away farket .
7	ask semidark of alış of floyd to feel shake server littl away farket .

Table 4.11. Translated English output sentence 3

Turkish sentence 4: Garion'un bu cevabı sonsuz acısını daha da artırmış gibi kasıldı Torak'ın yüzü.

English sentence 4: Torak's face twisted as if his perpetual agony had been doubled by Garion's refusal.

Trial ID	Processed Input Sentence-TR
5	garion bu cevap son acı daha da artır gibi kas torak yüz .
6	garion bu cevap son acı daha da artır gibi kas torak yüz .
7	garion bu cevap son acı daha da artır gibi kas torak yüz .

Table 4.12. Processed Turkish input sentence 4

Trial ID	Translated Output Sentence-EN
5	garion everyon get letter acı peasant ' artır : kas torak about .
6	garion everyon get the june acı our ' artır : kas torak about
7	garion everyon get the june acı our ' artır : kas torak about

Table 4.13. Translated English output sentence 4

5. CONCLUSION AND DISCUSSION

We worked on BOUN Corpus [1] and tried to divide it into smaller meaningful sections. After dividing and editing, we have decided to create a machine translation system based on the BOUN Corpus [1].

At the beginning of our project, we did a state-of-the-art research. We tried to find as many Turkish-English corpus as we can. Based on our research, we learned how a corpus should be and saw various methods to create a corpus. We also concluded that there wasn't an extensive Turkish-English corpus that is created aside from the corpus that is created by Boun-Itu [1].

After seeing that the Boun [1] corpus contains lots of mixed texts, we tried to divide the into their respective books. We did it in the way that is described in the Methods section.

We tried to roughly examine the corpus to see if it was aligned well during our separation work as well. We saw that most of the corpus was well aligned but it has some small mistakes at the beginning of some books and some misalignment about the bibliography of the books. We also saw that some of the books were put in the corpus twice, hence the corpus contained duplicate lines which are an unhealthy feature to have in a corpus.

At the end of the last semester we divided 200000 lines of the Boun corpus [1] and created a more organized corpus. This semester, we continued our project by creating a machine translation system that uses our corpus. In the first step, we decided to create a machine translation system using Moses [2]. We had several problems when we were trying to install and run Moses. It was especially hard to solve our problems because there weren't enough resources about Moses online. We spent lots of time trying to figure out the program.

After finding solutions to our problems, another hardship we faced was that running of the Moses [2] took a long time because there were too many steps to execute like tokenization, true-casing and cleaning the corpus. After that we needed to create a language model, tune it and test it. Those steps required the execution of many commands and the execution of the commands was long due to the size of the corpus. To make the process faster we decided to write a shell script. Because we both didn't do it before, we will first do some research on it.

After being able to run Moses [2] with no problem, we first tried whole corpus and a very small part of corpus to test out Bleu scores. As we expected, the Bleu score of the results increased with the increasing size of corpus that we used for training. When we were doing our first trials, we also saw that it took a really long time to run the Moses. Especially tuning stage took a really long time. This made our experiments significantly harder because the maximum number of experiments that we could conduct was one per day. Conducting experiments got harder in the next stages after Moses [2] stopped working on Sumeyra's computer. The main reason for this error was that Moses [2] requires a lot of memory space for running and Sumeyra didn't have this. We don't know really how we could speed up this process but it really wasn't practical.

As we decided to try different sizes of the corpus to see the results, we first tried different sizes. We tried whole corpus and a small proportion at first. Our result from this first trials was that indeed, the increasing training corpus size increased the scores. After contemplating about train, tune and test data proportions we decided to apply a 80%, 10%, 10% percentage for the data.

After completing the experiments about the corpus size we decided to pre-process the training data to see if we could improve the current results. We tried taking the stems of the words, removing the punctuation marks, removing stop-words and different combination of those pre-processings together for the data. After these trials, we discovered that we could get the best result with the stemmed words and 4-gram language model. In this step the main problem that we faced was the Bleu scores that we achieved. All of the Bleu scores that we achieved were less than 1 which can be

considered fairly bad. We didn't know what caused these bad scores as we applied every step of running Moses correctly as instructed in the documentation.

We also tried to install a neural machine translation as we decided at the beginning of our project but neural machine translation systems generally required high computing power and both Sumeyra and Irem worked with Ubuntu subsystem on Windows and didn't have enough resources to execute healthy experiments.

In this project, we learned how to prepare data for a machine translation system. We never prepared data before and we saw that it was a very hard and tedious job. But in every machine translation system, data is very important to get good results. Also processing data to get the maximum results is also very important. In conclusion, in this project we had the opportunity to try preparing data and process these data to create an optimum machine translation system.

6. FUTURE WORK

Up until now we prepared corpus data. We split the corpus into smaller parts and we separated them by the genre. After creating a reasonable corpus we experimented on best parameters and data processing algorithms to find the optimum translation results. These results can be found on Section 4.

The future work can be done on the Bleu scores of the statistical machine translation system. We compared the Bleu scores of different data for the training and concluded the best resulting one based on comparison. However, our Bleu scores have a problem that they are below one which shouldn't be like that. Normally, for statistical machine translation systems scores around 30 are normal. We didn't know the problem so one possible future work can be to inspect the problem about the scores.

Most important future work is the publishing the corpus that we prepared. The main corpus resource [1] we used is not a public one and it doesn't belong to us so we couldn't put it in a public Github repository. In the future, we are expecting the corpus that we created to be shared so everyone can work on it.

REFERENCES

1. Güngör, T., *BOUN-ITU Turkish-English Parallel Corpus*, 2019.
2. Philipp Koehn, A. B. C. C.-B. M. F. N. B. B. C. W. S. C. M. R. Z. C. D. O. B. A. C. E. H., Hieu Hoang, “Moses: Open Source Toolkit for Statistical Machine Translation”, *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.
3. “Moses”, <http://www.statmt.org/moses/>.
4. Christodoulopoulos, C., *Bible Corpus For Languages*, 2018, <https://github.com/christos-c/bible-corpus>.
5. Özbek, A., *Turkish-English Parallel Corpus*, 2019, <https://github.com/maidis/turkish-parallel-corpora>.
6. Ataman, D., “Bianet: A Parallel News Corpus in Turkish, Kurdish and English”, *LREC 2018 Workshop MLP-Moment*, 2018.
7. Biçiçi, E., *Turkish-English Parallel Corpus*, 2019, <https://github.com/bicici/SMTData>.
8. Sezer, T., “TS Corpus Project: An online Turkish Dictionary and TS DIY Corpus”, *European Journal of Language and Literature Studies*, 2017.
9. “The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2010”, <https://www.isca-speech.org/archive/iwslt10/papers/slta183.pdf>.
10. “Moses Installation For Developers”, <http://www.statmt.org/moses/?n=Development.GetSt>
11. “Moses Github Repository”, <https://github.com/moses-smt/mosesdecoder/tree/RELEASE>

12. “Moses”, <http://www.statmt.org/moses/?n=Moses.Baseline>.
13. “NLTK - PorterStemmer”, <https://www.nltk.org/api/nltk.stem.htmlmodule-nltk.stem>.
14. “PorterStemmer Algoritm”, <https://tartarus.org/martin/PorterStemmer/>.
15. “Zemberek WebSite”, <http://zembereknlp.blogspot.com/>.
16. “Zemberek Tutorial”, <https://medium.com/@melikebektas95/zemberek-k%C3%BCt%C3%BCph>.
17. “Nematus: a Toolkit for Neural Machine Translation”,
<https://arxiv.org/pdf/1703.04357.pdf>.
18. “Marian WebSite”, <https://marian-nmt.github.io/examples/mtm2017/intro/>.
19. Papineni, S. R. T. W., Kishore and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation”, *Research Report RC22176, IBM Research*, 2001.