transLogos 2020 Vol 3 Issue 1
Doğru, Gökhan, pp. 98–121
https://dx.doi.org/10.29228/transLogos.23

transLogos
A Translation Studies Journal

© Diye Global Communications
diye.com.tr | diye@diye.com.tr

Statistical Machine Translation Customization between Turkish and 11 Languages

Gökhan DOĞRU[*]

Statistical Machine Translation (SMT) has been the dominant corpus-based machine translation (MT) approach in the last twenty years. While SMT has been studied in detail among European languages, it has not been studied sufficiently in language pairs including Turkish as source or target language, and its study has been limited mostly to English ↔ Turkish language pair. This study aims to broaden the perspective on Turkish corpus-based MT studies by training MT engines between Turkish and a wide variety of languages with different features. It surveys customized SMT between Turkish and 11 different languages. Twenty-two SMT engines have been trained in KantanMT with open parallel corpora using Turkish as both source and target language. Three automatic evaluation metrics F-Measure, BLEU, and TER have been used for evaluating MT quality. Due to the variations in the corpus quality and size, highly varying results have been achieved. While Turkish ↔ Catalan engines have had the highest automatic evaluation scores, Turkish ↔ Arabic engines have had the lowest automatic scores. While the quality results are highly varying across languages, we obtain baseline scores for a wide variety of languages coupled with Turkish. These results may provide a reference point for evaluating future MT systems including Turkish.

Keywords: statistical machine translation customization; Turkish; automatic evaluation metrics; translation quality evaluation; parallel corpus

1. Introduction

Statistical Machine Translation (SMT) has been the dominant machine translation (MT) paradigm in the last two decades (Lumeras and Way 2017; Koehn 2009). The increasing amount of publicly available parallel corpora such as Europarl (Koehn, 2005) and the introduction of the free and open source MT customization toolkit Moses (Koehn et al. 2007) have accelerated the adoption of SMT both in academia and in the industry. While language pairs such as French ↔ English and Spanish ↔ Portuguese have been studied and high-quality results have been reported, MT studies on language pairs including Turkish have been quite limited. The available studies

[*] Predoctoral researcher at Universitat Autònoma de Barcelona.
E-mail: gokhan.dogru@uab.cat; ORCID ID: https://orcid.org/0000-0001-7141-2350.
(Received 23 April 2020; accepted 10 June 2020)

focus either on English ↔ Turkish MT (Oflazer and Durgar El-Kahlout 2007) or on MT between Turkic languages (Tantuğ and Adalı 2018). The study of Francis Morton Tyers and Murat Serdar Alperen (2010) is one of the few studies that reports results between Turkish and seven languages from the Balkans as well as English. In this preliminary study, we widen the scope of languages and report results for SMT between Turkish and 11 languages with a broader aim of understanding the performance of Turkish MT systems.

Turkish is one of the most widely used languages in the world. Ethnologue[1] reports that there are nearly 78.9 million Turkish speakers (according to its latest estimate in 2018). Besides, as of 2020, Turkish ranks as fourth most widely used content language on the web with a share of 2.9%, according to a study by W3Techs.[2] When Google announced its transition to neural machine translation (NMT), it included Turkish among the first seven languages (French, German, Spanish, Portuguese, Chinese, Japanese, Korean, and Turkish) to translate from and into English stating that "[t]hese [languages] represent the native languages of around one-third of the world's population, covering more than 35% of all Google Translate queries."[3] Considering all these facts as well as the increasing integration of Turkish people and industry into the global society and the accompanying demand for faster translation and communication, it is fundamental to study Turkish language in the context of MT. In this study, the focus will be on SMT and Turkish. Although SMT is a mature field which has been studied for more than 30 years and is the dominant MT system in translation industry, its full potential for Turkish into different languages / different languages into Turkish has not been studied sufficiently. One of the basic reasons for not discovering MT between Turkish and different languages has been the lack of publicly available parallel corpora used for training MT engines in the 2000s (Durgar El-Kahlout and Oflazer 2006). However, today large open parallel corpus collections are available on the OPUS corpus (Tiedemann 2012). Although corpora are not available for all the subject domains, there are enough number of parallel corpora

[1] "Turkish," *Ethnologue*, accessed June 11, 2018, https://www.ethnologue.com/language/tur.
[2] "Historical yearly trends in the usage statistics of content languages for websites," *W3Techs*, accessed April 22, 2020, https://w3techs.com/technologies/history_overview/content_language/ms/y.
[3] "Found in Translation: More Accurate, Fluent Sentences in Google Translate," *Google* (blog), published November 15, 2016, accessed November 6, 2018, https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

for different languages ↔ Turkish to train MT engines. Finally, while there is a growing tendency towards using NMT systems, some studies (e.g., Castilho et al. 2018) find in their automatic and human evaluation that SMT may still provide better quality results in some scenarios compared to NMT. From this, we can infer that studies on Turkish SMT can provide invaluable insights about when to select which MT paradigm.

Using a state-of-the-art SMT platform based on Moses and open parallel corpora, we have trained MT engines of Turkish ↔ 11 languages and obtained 22 engines in total, and subsequently we have evaluated the quality of these engines with three automatic quality metrics. The purpose of this preliminary study is to achieve baseline automatic MT evaluation scores for Turkish ↔ different languages MT engines and allow other researchers who study Turkish MT to compare their scores and see their relative improvements.

Section 2 explains the features of Turkish which make it a challenging target for SMT. Section 3 includes a detailed description of the corpus, tools, and methodology of the study. The automatic evaluation results for all language pairs have also been provided in this section. Section 4 discusses the results for 22 engines, while the last section, section 5, concludes the study.

2. The Challenging Features of Turkish for SMT

Different factors including the size and quality of parallel corpora, type of MT system, specificity of the translation domain as well as the grammar of each language involved, or grammatical similarity between the language pairs influence the quality of MT engines. Although it is hard to make generalization over which factor is more crucial, it is widely held that grammatical similarity between language pairs has a positive effect on the quality of MT engine in SMT systems. In a study where they compare automatic and manual evaluation of MT, Philipp Koehn and Christof Monz (2006, 109) state that "[i]t is well known that language pairs such as English-German pose more challenges to machine translation systems than language pairs such as French-English. Different sentence structure and rich target language morphology are two reasons for this." Hence, it is necessary to understand the grammatical features of Turkish relevant to SMT. This explication will allow us to see, at least partly, why there are differences of MT engine quality when the same or similar parallel corpora are used for training. In this part of the study, we will describe the features of Turkish which are relevant to SMT.

2.1 The Family of Turkish Language and Its Rich Morphology

Turkish has been a challenging language for SMT, and several studies have been made for solving the problems arising from the grammatical features of Turkish (Durgar El-Kahlout and Oflazer 2006; Oflazer and Durgar El-Kahlout 2007; Tantuğ, Oflazer, and Durgar El-Kahlout 2008). Turkish language (Istanbul Turkish or Anatolian Turkish) belongs to Ural-Altaic language family, under the subfamily of Turkic language family of Altaic language family. Turkic language family consists of 34 languages. A. Cüneyd Tantuğ and Eşref Adalı give a detailed account of the grammars of Turkic languages and explain:

> [a]ll Turkic languages have a very productive inflectional and derivational morphology where suffixes are affixed to a root word or to another suffix. While suffixes can be different among Turkic languages, the morphophonology and morphotactics rules are nearly same for all Turkic languages. (2018, 239–240)

For example, the Turkish word '*YAPAMAYACAKLARINDAN*' (meaning: BECAUSE THEY WILL NOT BE ABLE TO DO IT) has eight suffixes: YAP + A + MA + Y + ACAK + LAR + I + N + DAN, 'YAP' being the root. This creates the one-to-many and many-to-one alignment problems in the training phase of SMT systems. As in the previous example, the one word in Turkish (unigram) needs to be mapped to nine words in English (9-grams) for full correspondence. In the preprocessing before MT training, several studies (Oflazer and Durgar El-Kahlout 2007; Durgar El-Kahlout and Oflazer 2010; Bektaş et al. 2016) suggest using a morphological analyzer to analyze the root and morphemes in the Turkish corpus and make a morpheme alignment between Turkish and English sentences.

It is possible to argue that this problem will be less common between Turkic language pairs such as Turkish ↔ Turkmen and Azerbaijani ↔ Turkish, which have similar morphological structures.

2.2 Flexible Word Order

The standard word order in Turkish is Subject + Object + Verb (SOV). However, since the subjects, objects, and verbs are marked with suffixes that determine who does the action and on what object, it is possible to change word order without much change in sentence meaning. Let us take the sentence *İnek çiçeği yedi*. It means 'The cow ate the flower.' In Turkish, all combinations of word order will be equally grammatical and valid with very little change in meaning:

(i) İnek çiçeği yedi.
(ii) İnek yedi çiçeği.
(iii) Yedi çiçeği inek.
(iv) Yedi inek çiçeği.
(v) Çiçeği inek yedi.
(vi) Çiçeği yedi inek.

The word order flexibility creates a problem when Turkish is a target language in an SMT engine, and a language model (LM) is being trained for Turkish. Nevertheless, this problem may be observed less when standardized text types such as medical texts and academic texts are used for training corpora. It can be predicted that the use of a parallel corpus including subtitling texts may deteriorate the quality of the engine if no preprocessing is used since this type of text will include more spoken language with flexible sentence structure.

2.3 Lack of High-Quality Specific Domain Open Parallel Corpora for Turkish

MT training requires high volumes of parallel corpora. After the development of corpus-based MT systems, the need for high-quality parallel corpora has increased. And in the last two decades, international organizations such as the European Union (EU) and the United Nations (UN) have published their multilingual parallel corpora for free use in different file formats, including, most importantly, translation memory exchange format (TMX). This has helped researchers to experiment with these corpora and develop their systems to achieve better MT engines. Today parallel corpora from these two international organizations and from some other (mostly voluntary) translation projects including TED Talks and OpenSubtitles are compiled in OPUS Corpus Project

(Tiedemann and Nygaard 2004). OPUS Corpus Project has a huge database of translations for many language pairs, and it is very widely used as a source for free and open parallel and monolingual corpora. However, as we have underlined above, the quality of the parallel corpora is very significant in SMT. The "Garbage In, Garbage Out" motto is generally cited when the researchers want to state that if training corpora quality is bad, the quality of the MT engine will be bad as well. Before focusing on the problems with Turkish corpora in OPUS, we need to try to define what high-quality parallel corpora means.

Translation quality evaluation is still a highly debated topic, and there are interesting approaches such as the Multidimensional Quality Metrics (MQM)[4] to measure the quality of translation and annotate the types of translation errors. However, in our definition of high-quality translation, we do not mean having high scores under such a detailed evaluation. In a more modest sense, we consider a translation of high-quality when it is translated by a professional translator, reviewed by at least one reviewer, and published; to borrow the term used in MT Post-editing Guidelines of TAUS,[5] it should be of "publishable quality." Although the corpora of the EU and those of the UN meet this quality criterion, they are not available in Turkish simply because Turkish is not an official language of any of these institutions. Hence, large portions of the available parallel corpora from/into Turkish in OPUS are not professional translations but volunteer translations coming from the projects of TED Talks, OpenSubtitles, Global Voices, Wikipedia, etc. This does not necessarily mean that translations in these projects are of low-quality, but we believe that the involvement of professional translators is crucial for high-quality translation. The detailed description of the corpora used for each language pair will be provided in the following section. To sum up, the lack of high-quality free and open parallel corpora has hindered the development of Turkish MT, and research shall be conducted to create the necessary data for training Turkish MT engines.

[4] "Multidimensional Quality Metrics (MQM) Definition," *QT21 – Quality Translation 21*, last modified December 30, 2015, accessed December 3, 2018, http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

[5] "MT Post-editing Guidelines," *TAUS – The Language Data Network*, published November 2010, accessed December 3, 2018, https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines.

3. A Panorama of Turkish SMT with 11 Languages

SMT systems require training for (i) translation model (TM) which includes calculating the probabilistic patterns and building the bilingual phrase tables and (ii) language model (LM) which includes calculating the probabilistic target language grammar. TM is built through the parallel corpora while LM can be built either by the target side of the parallel corpora alone or by extra volumes of monolingual corpora. It is commonly agreed that unlike parallel corpora, LM can be as big as possible (Hearne and Way 2011).

In this section, we survey the quality of SMT between Turkish and 11 different languages in both directions. Reasons for selecting these languages are provided in the following subsection. This study will provide a baseline for comparison to other studies concentrating on different language pairs including Turkish.

3.1 The 11 Languages

We have selected 11 different languages: Arabic, Azerbaijani, Catalan, Chinese, English, French, German, Japanese, Persian, Russian, Spanish. While creating this collection of languages, our main criterion has been the availability of enough parallel corpora, which is the basis of an SMT engine. Secondly, we have tried to select the languages that are most commonly used in translation industry in Turkey. And finally, we have also tried to test the compatibility of Turkish with syntactically, semantically, and morphologically different (or similar) languages. Azerbaijani is selected for being a Turkic language which is closely related to Turkish and for testing the hypothesis that similar languages will yield better quality MT. Spanish, Catalan, and French are Romance languages; German and English are Germanic languages; Persian and Arabic share the same alphabet but have different grammars and are of different language families. Japanese and Turkish share a similar dependency representation in sentences (Tantuğ and Adalı 2018, 137) which may or may not influence quality, but the alphabet is different. Russian and Chinese have very different alphabets and grammars compared to Turkish. We have conducted the trainings in two directions to discover if the directionality affects MT quality.

Table 1. List of the 11 languages paired with Turkish

1	Arabic (ar)	↔	Turkish
2	Azerbaijani (az)	↔	Turkish
3	German (de)	↔	Turkish
4	Chinese (zh-cn)	↔	Turkish
5	French (fr)	↔	Turkish
6	English (en)	↔	Turkish
7	Spanish (es)	↔	Turkish
8	Japanese (ja)	↔	Turkish
9	Russian (ru)	↔	Turkish
10	Catalan (ca)	↔	Turkish
11	Persian (fa)	↔	Turkish

We will discuss the results for each language in the following section. Below we describe our tool and corpora.

3.2 The Tool and the Corpora

3.2.1 The Tool. We have used KantanMT[6] to train all the engines. KantanMT is a Moses-based, “cloud-based Statistical Machine Translation (SMT) platform.”[7] Being a cloud-based platform, it allows for training engines independent of computer platform. And having a user-friendly interface, it makes it possible to work without demanding technical skills. Most importantly, it has in-built automatic evaluation metrics, which is very helpful for reporting and following the qualities of the engines. Hence, since it is a stable platform, it has been possible to concentrate on the parallel corpora preparation and results without making any change about the core of the system.

[6] Available at https://kantanmt.com/.

[7] “Moses Use Case: KantanMT.com,” *KantanMT Blog*, published January 21, 2015, https://kantanmtblog.com/2015/01/21/moses-use-case-kantanmt-com/.

3.2.2 The Corpora. We have used the corpora available in the OPUS corpus nearly for all the engines. The only exception has been the Turkish ↔ Azerbaijani language pair since enough training data for it has not been available. We have needed to create new parallel corpora out of publicly available documents such as the Azerbaijani Constitution. For the remaining language pairs, we have used the following corpora: GNOME, Tanzil, Tatoeba, KDE4, Ubuntu, OpenSubtitles,[8] PHP, GlobalVoices, KDEdoc. We have aimed to have similar number of source words in each engine. However, it has not always been possible due to the available sizes of the corpora and the characteristics of the languages involved. Since each corpus is from a different domain (news, subtitle, religion, localization, etc.), the created engines are of mixed domains. It should also be highlighted that in most of the multilingual corpora, normally translations are performed from English into different target languages. In other words, a Spanish–Turkish corpus is usually derived from an English–Spanish and an English–Turkish corpus, which, in turn, may affect the quality of the resulting MT engine. Below we will describe the text type of each of the corpora we have benefited from.

(i) The GNOME corpus: This corpus is hosted in the OPUS corpus database. It includes the localization strings from the GNOME Project. The translation of the project is explained as follows:

> The bulk of GNOME translations are performed by native speakers on a volunteer basis. They take sentences in the original English, supply the appropriate translation, and add the file containing this information to the GNOME Git repository so that the next release of the software contains the new language.[9]

(ii) The Tanzil corpus: The Tanzil corpus includes translations of Quran from the Tanzil Project:

> Tanzil is a Quranic project launched in early 2007 to produce a highly verified Unicode Quran text to be used in Quranic websites and applications. Our mission in the Tanzil

[8] We have not used the OpenSubtitles 2016 or 2018 versions. They are not available for our every language pair, and when available, their size is too big to handle in the tool.

[9] "Localising GNOME Applications: So You Want to Translate GNOME?" *GNOME Wiki*, accessed December 11, 2018, https://wiki.gnome.org/TranslationProject/LocalisationGuide.

> project is to produce a standard Unicode Quran text and serve as a reliable source for this standard text on the web.[10]

The sentence strings include religious content, and the quality of translation and alignment is generally high.

(iii) The Tatoeba corpus: The Tatoeba corpus includes translations of short, generic sentences by volunteers. It is "a large database of sentences and translations. Its content is ever-growing and results from the voluntary contributions of thousands of members."[11]

(iv) The KDE4 corpus: The KDE4 corpus is a relatively small corpus, and it includes short localization strings translated by volunteers.[12]

(v) The Ubuntu corpus: This corpus includes localization strings from Ubuntu translated by a community of volunteers.[13] The sentences are generally short, and some of them include placeholders which can create problems during the training of MT systems if not cleaned manually or automatically.

(vi) The OpenSubtitles corpus: OpenSubtitles is a very large corpus from a subtitle website called OpenSubtitles. It contains subtitle strings from movies, documentaries, TV shows, etc. Translations are compiled from the Internet and/or translated by volunteers/fans.[14]

(vii) The PHP corpus: PHP is a server-side scripting language, and the corpus includes localization strings. In the OPUS corpus, it is stated that "[t]he corpus is rather noisy and may include parts from the English original in some of the translations."[15]

(viii) The GlobalVoices corpus: The GlobalVoices corpus includes news stories from all over the world. The translations are made by volunteers through the translation tool of the website.[16]

[10] "Tanzil Project," *Tanzil*, accessed January 9, 2019, http://tanzil.net/docs/Tanzil_Project.
[11] "What is Tatoeba?" *Tatoeba*, accessed January 9, 2019, https://tatoeba.org/eng/about.
[12] *KDE.org*, accessed January 9, 2019, https://www.kde.org.
[13] "Translations," *Ubuntu*, accessed January 9, 2019, https://translations.launchpad.net/ubuntu.
[14] *OpenSubtitles.org*, accessed January 9, 2019, https://www.opensubtitles.org.
[15] "PHP," *OPUS – An Open Source Parallel Corpus*, accessed January 9, 2019, http://opus.nlpl.eu/PHP.php.
[16] "Translators Guide," *Global Voices Community Blog*, accessed January 9, 2019, https://community.globalvoices.org/guide/lingua-guides/lingua-translators-guide/.

(ix) The KDEdoc corpus: The KDEdoc[17] corpus includes the translations of KDE manuals by volunteers. It is a relatively small corpus, and the manuals are IT-oriented.

After this review of the corpora, it can be observed that most of them are translations by volunteers, and the domains are localization (five corpora), religion (one corpus), subtitling (one corpus), general (one corpus), and news (one corpus). Important domains such as medicine, law, and politics are missing to make an engine as comprehensive as possible.

3.3 Methodology

The study includes three phases: (i) corpus preparation, (ii) training, (iii) quality evaluation. In the first phase, available corpora in the form of TMX have been collected for each language pair (mostly obtained from the OPUS corpus). Note that since parallel corpora have been readily available, methods such as web crawling, alignment, and cleaning for corpus preparation have not been performed. Each corpus has been trained in two directions (for example, ES → TR, TR → ES) in KantanMT. No separate extra monolingual corpus has been added, and only the target side of the parallel corpus has been utilized for the LM in each engine. Although it is possible to add monolingual files and glossary/terminology files to KantanMT, we have only added translation memories for the purpose of the study. And we have not used the stock data of KantanMT, either. There are no strict word count limits for MT training. However, KantanMT has some suggested word count types for bilingual source language word count (WC), unique word count (UWC), and monolingual word count (MWC) to achieve a good quality engine: five million words for WC, three hundred thousand words for UWC in the case of specific domain engine and five hundred thousand words for UWC in the case of general domain engine, and finally two to three million in-domain words for monolingual data. It has not been possible to reach these thresholds in all our engines due to corpus quantity constraints.

[17] "KDEdoc," *OPUS – An Open Source Parallel Corpus*, accessed January 9, 2019, http://opus.nlpl.eu/KDEdoc-v1.php.

When training is completed, KantanMT provides three automatic evaluation metrics: F-Measure, BLEU, and TER. According to KantanMT,[18] F-Measure is an automatic calculation of recall and precision, and the more words the engine correctly selects, the better the vocabulary of the engine. The aim should be to achieve higher scores (a probable threshold of 70% indicates a good engine).[19] BLEU (Papineni et al. 2002), in KantanMT's view, is a calculation of fluency, and it measures phrase selection capability of the engine. Again, the aim should be a higher score (a probable threshold of 60% indicates a good engine). Finally, TER is an automatic calculation of the post-editing effort, and the score shall be as low as possible (aim for a score of 40% or lower). These three automatic evaluation metrics are used to compare the qualities of each engine in this study, which does not use human evaluation. A future study with both automatic and human evaluation can be very beneficial for Turkish MT studies.

3.4 Results

In this subsection, results for MT engines from different languages into Turkish are presented with F-Measure, BLEU, and TER scores first. Then, results for Turkish as the source language and other languages as the target are provided.

3.4.1 Eleven Languages to Turkish. There has been significant variation in the number of WCs and UWCs, which, in turn, has resulted in variations in the automatic quality scores of F-Measure, BLEU, and TER. The variation in corpus size has resulted from the lack of parallel corpora for the relevant language pair. Table 2 shows all the results together sorted by the highest number of WCs. The average quality scores of automatic evaluation metrics for the 11 engines are as follows: F-Measure 35.18%, BLEU 32%, and TER 87%. In KantanMT's criteria, an engine with these averages has "below average knowledge of your target domain and language" according to F-Measure score,[20] is "below average and will not produce highly fluent translations" according to

[18] "Understanding KantanBuildAnalytics Scores," YouTube video, 1:34, posted by "KantanMT," August 11, 2017, https://www.youtube.com/watch?time_continue=34&v=kIWrH9O-p6U.

[19] The score threshold recommendations for F-Measure, BLEU, and TER are provided in KantanMT's MT training course: KantanAcademy™, which is available within kantanmt.com.

[20] "F-Measure in BuildAnalytics," *KantanMT.com*, accessed June 3, 2020, https://kantanmt.zendesk.com/hc/en-us/articles/204656689-F-Measure-in-BuildAnalytics.

BLEU score,[21] and "will require a high level of post-editing" according to TER score.[22] From table 2, it can be observed that while the Persian → Turkish MT engine has a significantly bigger corpus (with a WC of 14,671,648), it has the lowest BLEU score. This fact implies that in the case of a language pair including Turkish, higher WC does not necessarily mean higher MT quality. And, the Catalan → Turkish engine has the highest UWC and the best F-Measure, BLEU, and TER scores. Although the CA → TR and EN → TR engines have very close WC and UWC, their F-Measure (64% vs. 43%), BLEU (64% vs. 39%), and TER (50% vs. 76%) scores are significantly different. This fact implies that factors such as corpus quality and linguistic factors also influence the quality of the engines.

Table 2. The WC, UWC, and F-Measure, BLEU, and TER automatic evaluation metric results of all the into-Turkish MT engines[23]

Engine Name	**Source**	**Target**	**WC**	**UWC**	**F-Measure**	**BLEU**	**TER**
Fa-Tr-Engine-1	fa	tr	14,671,648	132,752	28%	11%	91%
De-Tr-Engine-1	de	tr	6,031,100	141,753	32%	15%	91%
En-Tr-Engine-1	en	tr	5,993,472	200,963	43%	39%	76%
Ca-Tr-Engine-1	ca	tr	5,556,181	202,500	**64%**	**64%**	**50%**
Ar-Tr-Engine-1	ar	tr	3,966,320	102,354	19%	25%	108%[24]
Az-Tr-Engine-1	az	tr	3,091,177	33,232	31%	20%	111%
Es-Tr-Engine-1	es	tr	2,484,247	118,359	37%	41%	84%
Fr-Tr-Engine-1	fr	tr	2,464,426	117,811	24%	33%	98%
Ru-Tr-Engine-1	ru	tr	585,262	96,541	38%	42%	83%

[21] "BLEU in BuildAnalytics," *KantanMT.com*, accessed June 3, 2020, https://kantanmt.zendesk.com/hc/en-us/articles/205355285-BLEU-in-BuildAnalytics.

[22] "TER in Kantan BuildAnalytics," *KantanMT.com*, accessed June 3, 2020, https://kantanmt.zendesk.com/hc/en-us/articles/204658269-TER-in-Kantan-BuildAnalytics.

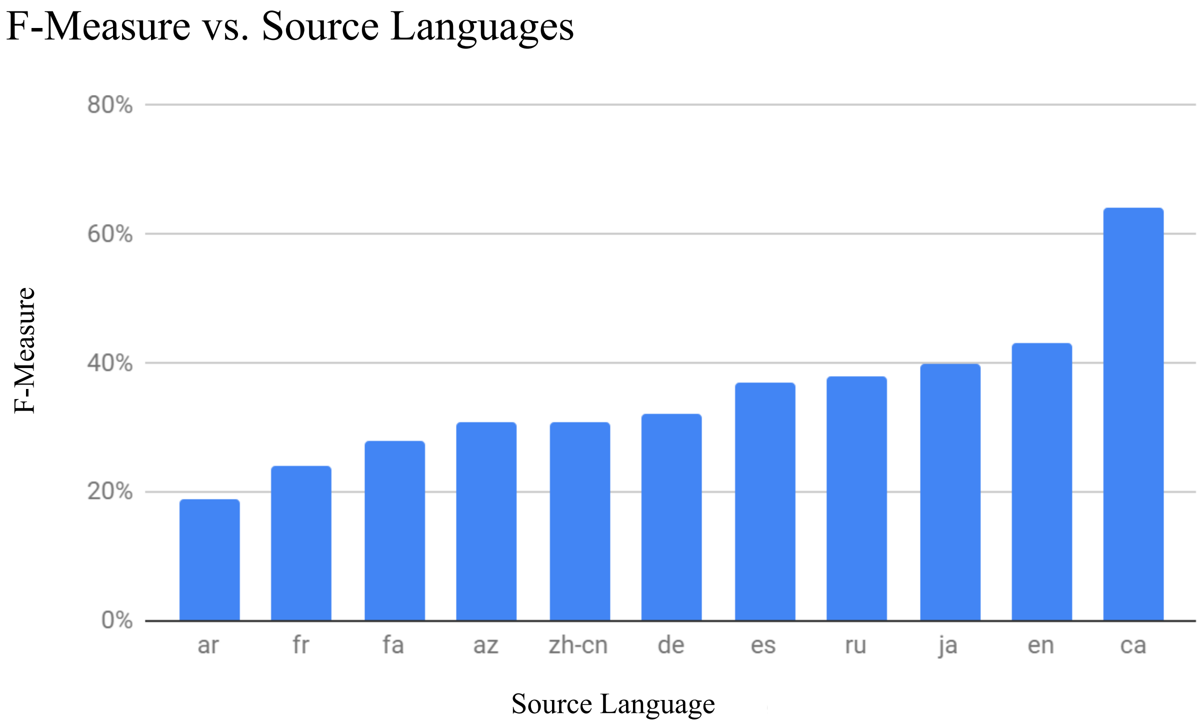
[23] Best scores are highlighted in bold.

[24] Upon our correspondence with the technical support of KantanMT, they have mentioned the reason why there is a score above 100% as follows: "Technically by definition TER does not have a maximum limit (in fact, using the percentage is a bit misleading in this example). TER is defined as the ratio between the number of edits (i.e., additions, deletions, and substitutions) and the number of words in the reference translation. If the edits happen to be more than the words in the reference translation, the TER ratio will be greater than one, and therefore the percentage score will be above 100%."

Ja-Tr-Engine-1	ja	tr	232,865	76,333	40%	42%	74%	
Zh-Tr-Engine-1	zh-cn	tr	127,436	16,051	31%	16%	93%	

3.4.1.1 F-Measure. The F-Measure scores are distributed between 19% and 64% (fig. 1). The highest F-Measure score has been obtained in the Catalan → Turkish engine while the lowest one in the Arabic → Turkish engine. Contrary to our assumption that Azerbaijani will have the highest score among the engines, we have observed that the score of the Azerbaijani → Turkish engine is one of the lowest. One of the reasons may be the low amount of parallel corpus. However, while Chinese has a smaller corpus size, its F-Measure score seems to be higher than that of Azerbaijani.

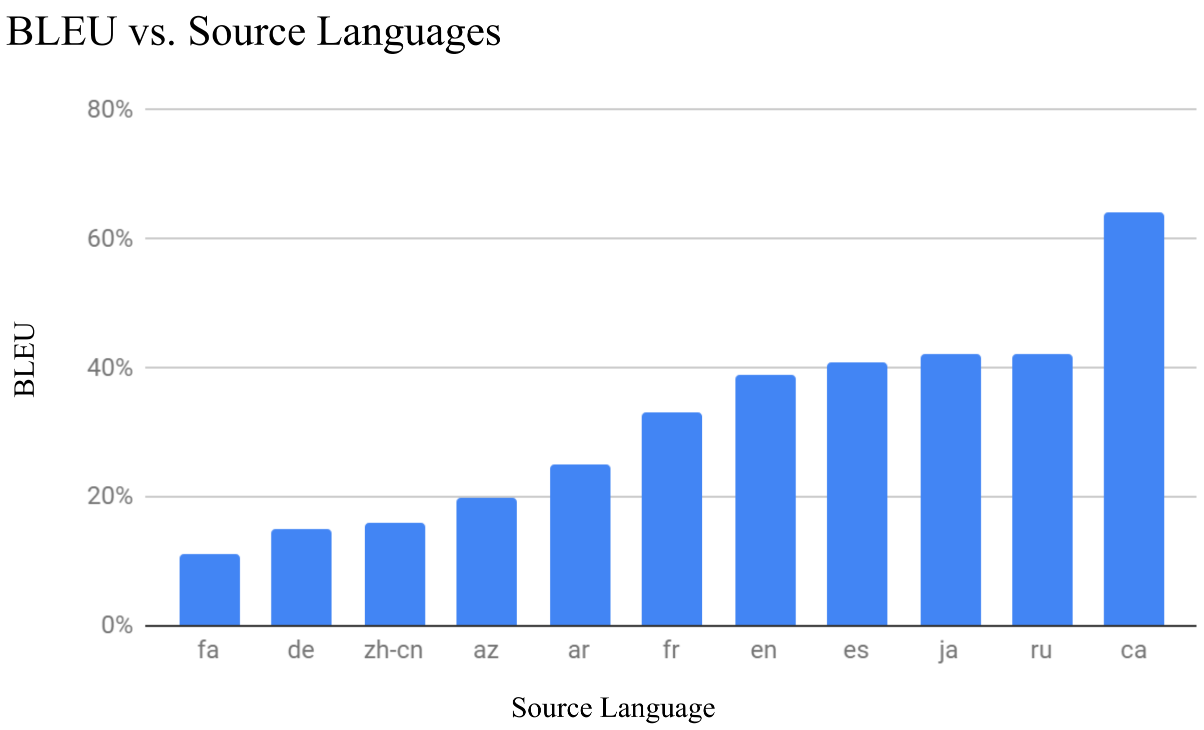
Figure 1. F-Measure scores for 11 engines as calculated by KantanMT



3.4.1.2 BLEU. The BLEU scores are distributed between 11% and 64% (fig. 2). The highest BLEU score has been obtained in the Catalan → Turkish engine while the lowest one in the Persian → Turkish engine. As in the case of F-Measure, contrary to our assumption that the Azerbaijani engine will have the highest score among the engines, we have observed that the score of the Azerbaijani

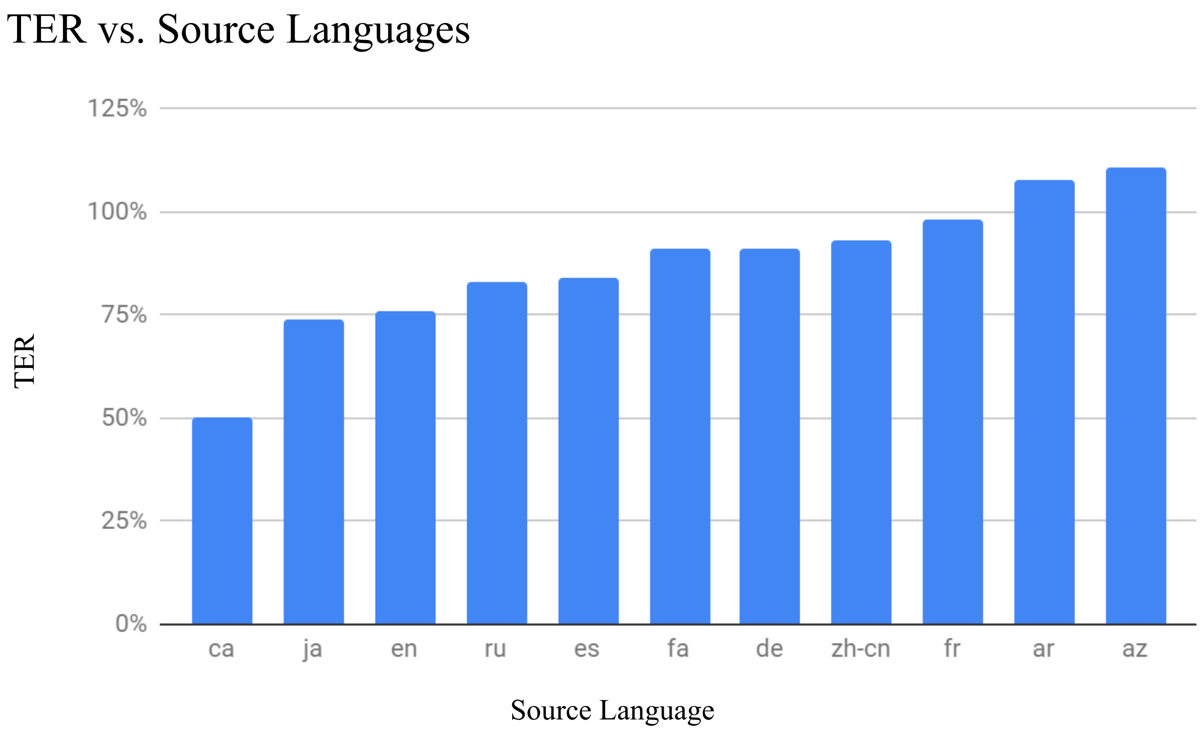
→ Turkish engine (20%) is again one of the lowest. The English → Turkish engine has a score of 39%, and it is just below the threshold of 40%, which means it will "produce reasonably fluent translations" according to KantanMT. Only Spanish, Japanese, Russian, and Catalan are above this threshold.

Figure 2. BLEU scores vs. source languages



3.4.1.3 TER. The lower the score of TER, the less the number of errors. TER score is mostly consistent with BLEU score in terms of the first five languages (fig. 3). However, the language that needs the highest amount of post-editing according to this score is Azerbaijani. Again, the Catalan → Turkish engine needs the least amount of post-editing.

Figure 3. TER scores vs. source languages



3.4.2 Turkish to 11 Languages. SMT engines having Turkish as the source language and different languages as the target language are evaluated in this subsection. The same corpora in each language pair have been reversed; Turkish has become the source language; and the trainings have been made accordingly. LM (the probabilistic target language grammar) was the same for all the engines in the 11 languages to Turkish setting. In other words, Turkish LM was used in each of them. However, in this Turkish to 11 languages scenario, LMs are different in each language pair. For example, in the Turkish to Spanish engine, the LM that will be trained and implemented will be (probabilistic) Spanish grammar, and in the case of Turkish to Azerbaijani, it will be the Azerbaijani LM.

When it comes to the automatic quality of each engine, the average F-Measure score is 44%, while the average BLEU score is 37%, and average TER score is 95%. These results show that MT from Turkish to other languages is better in terms of BLEU and F-Measure but requires more post-editing since the TER score is worse. Compared to the previous scenario (all languages

→ Turkish), only the engines with Japanese and Chinese languages have more WCs. And concerning UWC, all the engines have more unique words except for Russian, which has slightly less unique words compared to the first scenario. For example, in the first scenario, the Persian → Turkish MT engine had 132,752 unique words. The Turkish to Persian MT engine includes 157,170 unique words. The reason why there are more unique source words in this scenario can be explained by the morphologically rich nature of Turkish. In other words, due to the derivational suffixes in Turkish, there is a more frequent diversity of word forms. The verb form '*yapmak*' (to do) is present in the corpus as '*yaptı*' (s/he did), '*yaptılar*' (they did), '*yapacaklar*' (they will do), '*yapıyorlar*' (they are doing), '*yaptım*' (I did), etc. Hence, the MT quality in this scenario is expected to be significantly different from that of the first scenario in most of the languages.

Table 3. Turkish to 11 languages SMT engines trained in KantanMT with WC, UWC, and F-Measure, BLEU and TER scores provided[25]

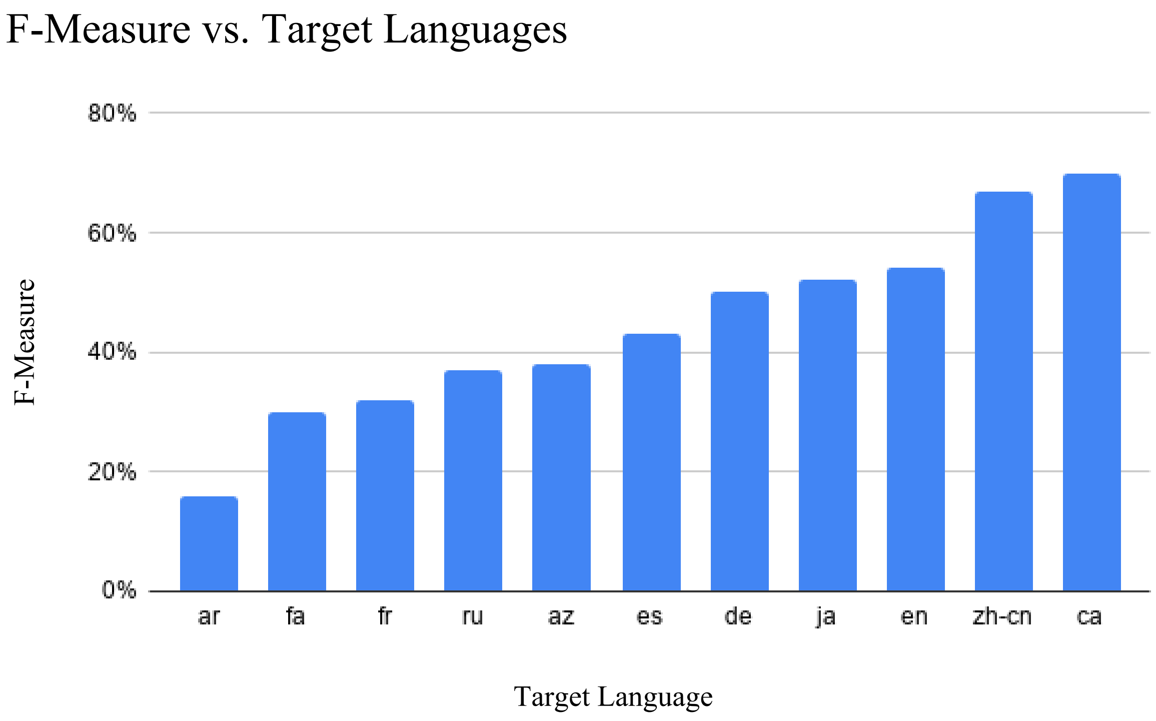
Engine Name	**Source**	**Target**	**WC**	**UWC**	**F-Measure**	**BLEU**	**TER**
Tr-Fa-Engine-1	tr	fa	10,094,431	157,170	30%	23%	147%
Tr-En-Engine-1	tr	en	4,888,629	445,982	54%	43%	70%
Tr-De-Engine-1	tr	de	4,684,887	210,322	50%	25%	70%
Tr-Ca-Engine-1	tr	ca	4,208,929	300,162	**70%**	**67%**	45**%**
Tr-Ar-Engine-1	tr	ar	2,753,150	127,967	16%	10%	218%
Tr-Az-Engine-1	tr	az	2,716,128	73,068	38%	32%	84%
Tr-Zh-Engine-1	tr	zh-cn	2,015,200	111,312	67%	55%	51%
Tr-Es-Engine-1	tr	es	2,003,395	235,901	43%	38%	92%
Tr-Fr-Engine-1	tr	fr	1,866,737	230,734	32%	34%	98%

[25] Best scores are highlighted in bold.

Tr-Ja-Engine-1	tr	ja	1,026,812	142,777	52%	40%	88%
Tr-Ru-Engine-1	tr	ru	566,618	90,589	37%	41%	84%

3.4.2.1 F-Measure. The Turkish → Catalan engine has the highest F-Measure with a score of 70% (fig. 4). Of all the 22 engines in this study, this is also the engine with the highest scores. Surprisingly, the Turkish → Chinese engine ranks as the second-best engine in terms of F-Measure, which is an unexpected result. We have conducted the training again to verify and have reached the same result. On the other side of the spectrum, the Turkish → Arabic engine has the lowest F-Measure score just as the Arabic → Turkish engine.

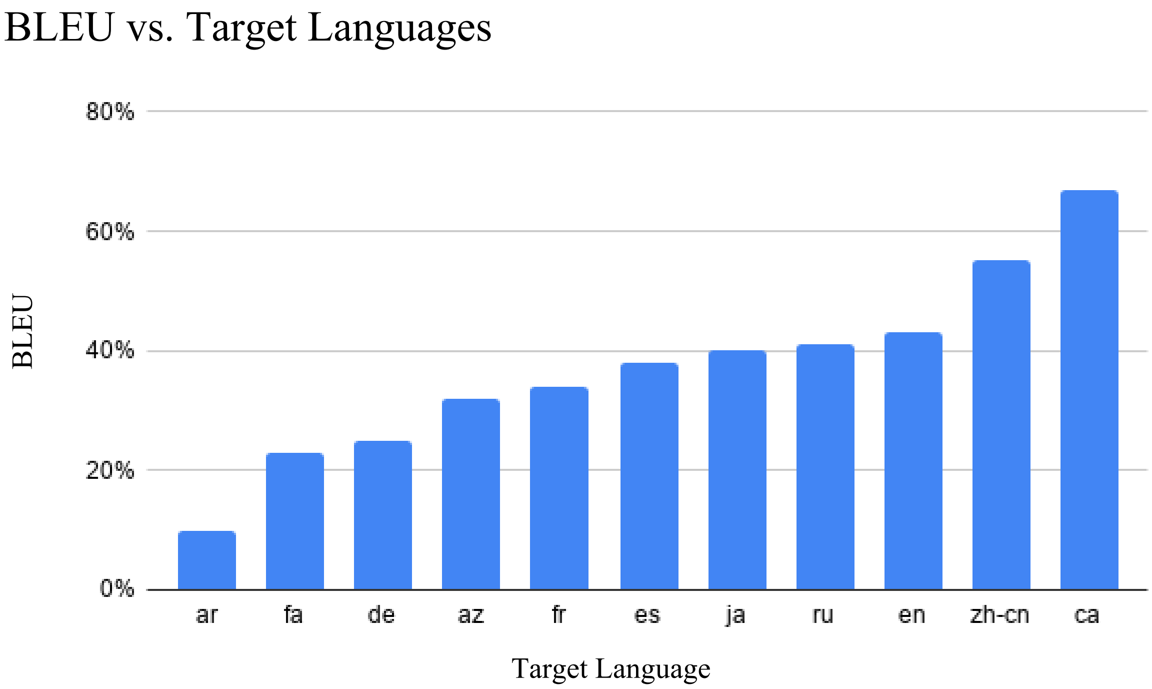
Figure 4. F-Measure scores for the Turkish to different languages engines illustrated by target language



3.4.2.2 BLEU. The Turkish → Catalan, Chinese, and English engines occupy the first three ranks in the BLEU scores (fig. 5). The difference that is worth mentioning in this figure is that of German.

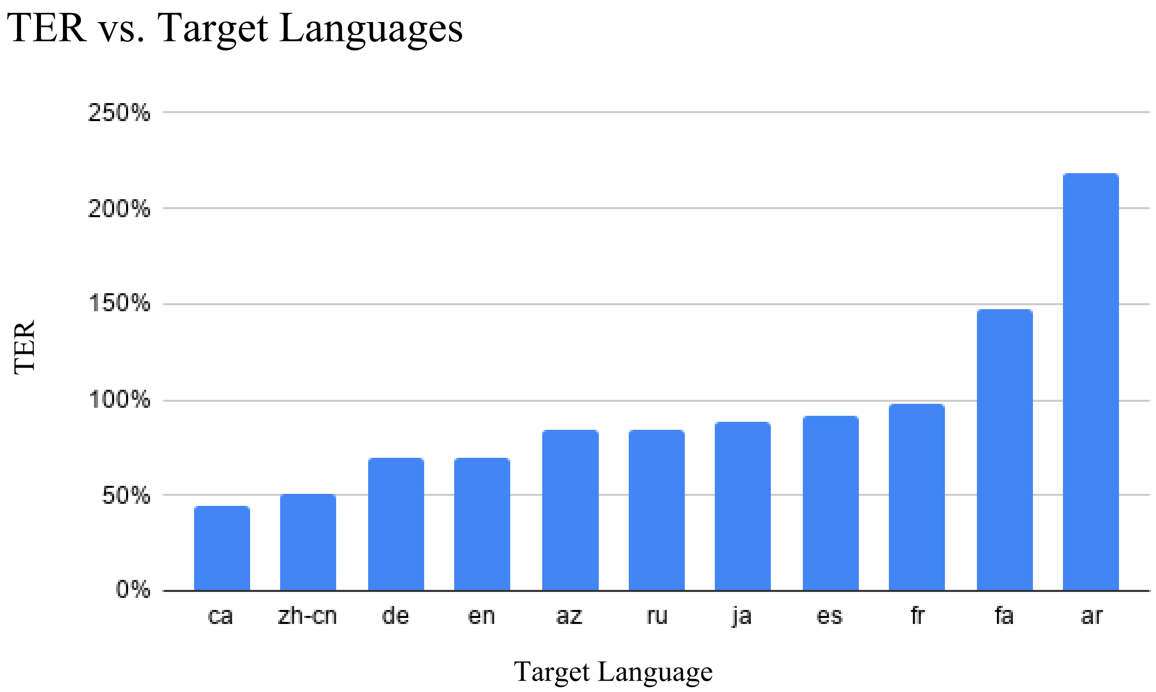
While the German engine is the fifth best engine in terms of F-Measure with a score of 50%, it has one of the lowest BLEU scores: 25%.

Figure 5. BLEU scores for the Turkish to different languages engines illustrated by target language



3.4.2.3 TER. The Turkish to Catalan engine has the best TER score (45%) among the 22 engines trained (fig. 6). With a 70% score of TER, German is the third-best engine, which is, again, a surprising finding as compared to its BLEU score.

Figure 6. TER scores for the Turkish to different languages engines illustrated by target language



4. Discussion

Our study has yielded interesting findings with 22 engines having Turkish as a source language and a target language. We have had a chance to observe the effect of three parameters, including parallel corpus size, similarity between languages, and UWC. As shown by the Persian ↔ Turkish engines, although a corpus has a sufficient size, this may not guarantee a better MT quality. Secondly, the Turkish → Azerbaijani engine has demonstrated that similarity between languages may only be an advantage if the corpus size is sufficiently big, which has not been the case in our experiment. Lastly, our comparison of automatic evaluation metrics with UWCs has not found a correlation between UWC and improvements in quality.

Automatic quality evaluation metrics have given us highly varying scores for 22 engines. Our training setting has been limited by several factors. The corpora used for training the engines have not been homogenous enough. While some language pairs have had a large amount of open parallel corpora, some have had very limited data. Besides, the translation quality of these corpora has not been very high, as explained previously. Another limitation is that the evaluations have only been made automatically. For more reliable results, human evaluation shall be conducted for each engine. Especially the high scores of Turkish → Chinese and Catalan ↔ Turkish and the low scores of Azerbaijani → Turkish are interesting phenomena that require further investigation to understand their outstanding results.

The Catalan → Turkish engine's being the engine with the highest score is an unexpected finding in this experiment. To confirm this finding, we have retrained the Catalan → Turkish corpora two times more. However, the results have not changed significantly. The BLEU scores have been as follows: 67% and 65%.[26]

The MT quality in each engine has been below the thresholds of KantanMT. In a real-world industrial production level, several preprocessing and postprocessing steps are needed, as explained by Kemal Oflazer and İlknur Durgar El-Kahlout (2007), for creating an engine that can yield accurate and fluent translations. Especially the size and quality of the training corpora are particularly important. The fact that the translations in the training corpora are performed by volunteer translators, aligned automatically (Tiedemann 2012), and are mostly indirect translations (e.g., Spanish–Turkish translations are derived from English–Spanish and English–Turkish translations) decreases the quality of the corpora. Besides, as A. Cüneyd Tantuğ, Kemal Oflazer, and İlknur Durgar El-Kahlout (2008) observe, the standard form of automatic evaluation metrics sometimes may not reflect accurately the quality of an MT system with Turkish, especially in the case of BLEU because of the agglutinative nature of Turkish. They suggest a custom BLEU score which they call BLEU+. As a whole, automatic evaluation metrics need to be accompanied by human evaluation studies to provide reliable information on the overall quality.

5. Conclusion

Turkish MT studies are comparatively very new, and there are still very few studies made on corpus-based MT in different languages paired with Turkish other than the English ↔ Turkish

[26] The BLEU score may change slightly because after each training different reference test sentences are randomly selected from the corpus.

language pair. For such studies, free and open high-quality parallel corpora are a prerequisite. Although thanks to the OPUS Corpus Project, it is now possible to experiment with different languages paired with Turkish, these corpora are pivoted from English. It will be interesting to have a parallel corpus directly translated between the involved languages. In short, the preparation of free and open high-quality parallel corpora between Turkish and other languages can lead to more studies on Turkish MT and to better MT quality results. This may also enable creating domain-specific engines. It should be noted that these corpora can also be used later for NMT, which needs even more parallel corpora for training.

We hope that the SMT engines that we have created can serve as a baseline for new MT studies and allow for new advancements.

Acknowledgments

I would like to thank Anna Aguilar-Amat and Adrià Martín-Mor for their valuable comments and recommendations. I am also grateful for the valuable comments of the two reviewers of my article. This research is supported and funded by AGAUR-FI 2017.

References

Bektaş, Emre, Ertuğrul Yılmaz, Coşkun Mermer, and İlknur Durgar El-Kahlout. 2016. "TÜBİTAK SMT System Submissions for WMT 2016." In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 246–251. doi:10.18653/v1/W16-2305.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. "Evaluating MT for Massive Open Online Courses: A Multifaceted Comparison between PBSMT and NMT Systems." *Machine Translation*, no. 32, 255–278. doi:10.1007/s10590-018-9221-y.

Durgar El-Kahlout, İlknur, and Kemal Oflazer. 2006. "Initial Explorations in English to Turkish Statistical Machine Translation." In *HLT-NAACL 06 Statistical Machine Translation Proceedings of the Workshop, 8–9 June 2006, New York City, USA*, 7–14. Madison, WI: Omnipress. https://www.aclweb.org/anthology/W06-3102.pdf.

———. 2010. "Exploiting Morphology and Local Word Reordering in English to Turkish Phrase-Based Statistical Machine Translation." *IEEE Transactions on Audio, Speech and Language Processing* 18 (6): 1313–1322. doi:10.1109/TASL.2009.2033321.

Hearne, Mary, and Andy Way. 2011. "Statistical Machine Translation: A Guide for Linguists and Translators." *Language and Linguistics Compass* 5 (5): 205–226. doi:10.1111/j.1749-818X.2011.00274.x.

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *The Tenth Machine Translation Summit: Proceedings of Conference, September 12–16, 2005, Phuket, Thailand*, 79–86. http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

———. 2009. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp, and Christof Monz. 2006. "Manual and Automatic Evaluation of Machine Translation between European Languages." In *HLT-NAACL 06 Statistical Machine Translation: Proceedings of the Workshop, 8–9 June 2006, New York City, USA*, 102–121. Madison, WI: Omnipress. http://www.statmt.org/wmt06/proceedings/pdf/WMT14.pdf.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." In *ACL 2007: Proceedings of the Interactive Poster and Demonstration Sessions, June 25–27, 2007, Prague, Czech Republic*, 177–180. Madison, WI: Omnipress. https://www.aclweb.org/anthology/P07-2045.pdf.

Lumeras, Maite Aragonés, and Andy Way. 2017. "On the Complementarity between Human Translators and Machine Translation." *HERMES*, no. 56, 21–42. doi:10.7146/hjlcb.v0i56.97200.

Oflazer, Kemal, and İlknur Durgar El-Kahlout. 2007. "Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation." In *ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic*, 25–32. Madison, WI: Omnipress. https://www.aclweb.org/anthology/W07-0704.pdf.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: A Method for Automatic Evaluation of Machine Translation." In *ACL 2002: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002, Philadelphia, USA*, 311–318. Association for Computational Linguistics. doi:10.3115/1073083.1073135.

Tantuğ, A. Cüneyd, and Eşref Adalı. 2018. "Machine Translation between Turkic Languages." In *Turkish Natural Language Processing*, edited by Kemal Oflazer and Murat Saraçlar, 237–254. Cham, Switzerland: Springer International.

Tantuğ, A. Cüneyd, Kemal Oflazer, and İlknur Durgar El-Kahlout. 2008. "BLEU+: A Tool for Fine-Grained BLEU Computation." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 1493–1499. http://www.lrec-conf.org/proceedings/lrec2008/pdf/382_paper.pdf.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Tiedemann, Jörg, and Lars Nygaard. 2004. "The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 1183–1186. http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf.

Tyers, Francis Morton, and Murat Serdar Alperen. 2010. "South-East European Times: A Parallel Corpus of Balkan Languages." In *Proceedings of the LREC'10 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, 49–53. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W22.pdf.