Noname manuscript No.
(will be inserted by the editor)

Systematic Review of Available corpora of low-resource Bangla Language for Machine Translation

Masum Ahmed · Shamihul Islam Khan Limon · Md Zobaer Hossain · Mohammad Abdullah Al Mumin · M Jahirul Islam

the date of receipt and acceptance should be inserted later

Abstract Machine Translation translates texts of one natural language into texts of another automatically. State of the art MT approaches use parallel corpora as their training data. This study aims at finding available Bangla ↔ English parallel corpora as well as measuring quality of them by using common linguistics features, evaluating well-known metrics by following statistical machine translation techniques. This paper reviews of existing publicly available Bangla ↔ English parallel corpora till 2020. Thirteen out of fourteen corpora have been founded publicly available. Most of the corpora were constructed using open source domain and pre-processed of their own way.

Keywords Machine Translation · Low-resource · Available Corpora · Bangla to English · English to Bangla · Statistical Machine Translation

Masum Ahmed
E-mail: *masumahmedeesha@gmail.com*

Shamihul Islam Khan Limon
E-mail: *shamihul78@student.sust.edu*

Md Zobaer Hossain
E-mail: *zobaer37@student.sust.edu*

Mohammad Abdullah Al Mumin
E-mail: *mumin-cse@sust.edu*

M Jahirul Islam
E-mail: *jahir-cse@sust.edu*

Shahjalal University of Science & Technology, Sylhet

Extended author information available on the last page of the article

1 Introduction

In the year 2012, the Encyclopedia Britannica declared that they will no longer print any publications. 244 years old publisher has moved out from the printing to publishing online as the technological revolution made access to the internet so easy [1]. This news is eight years old, recent statistics are more astounding. The data that has been created in the last two years is more than any previous human history. The exponential growth of the internet has made this possible. Internet subscribers of Bangladesh have reached 93.702 million users in 2019 at an increased rate of 15-16% per year [2]. This all proves how we have engaged ourselves in these modern technologies.

According to Common Sense Advisory, 72.1% of internet users prefer to visit sites in their native language [3]. But it is not feasible to use manual translations on billions of webpages or contents. Without any automatic translations system it is not possible to make the internet easier. Machine Translations(MT) is introduced to solve this problem. MT automatically converts the text of one language to another. Statistical machine translations(SMT) and Neural machine translations(NMT) are the two very common approaches that have been used in MT. Here both SMT and NMT are data-driven approaches. That means both of the techniques require corpora in order to develop an automatic translations system.

With approximately 228 million native speakers and another 37 million as second language speakers, Bangla

[1] *www.cmswire.com/customer − experience/why − machine − translation − matters − in − the − modern − era*
[2] *en.wikipedia.org/wiki/Internet*$_{in}$*Bangladesh*
[3] *omniscien.com/?faqs = why − do − i − need − machine − translation*

is the fifth most-spoken native language and the seventh most spoken language by the total number of speakers in the world [4]. But it is also true that Bangla is a low-resourced language in terms of machine translations. The number of studies and required resources related to Bangla MT is very low.

Since corpora are pre-requisite of MT systems, a large scale corpora will enhance the MT system in Bangla language. However, there are some publicly available corpora on Bangla ↔ English. But to the best of our knowledge there are no systematic analysis on these available parallel corpora. An analysis on corpora construction along with their data distribution and quality comparison will create an opportunity for the researchers on this field.

In this paper, we reviewed all the available Bangla ↔ English parallel corpora. While reviewing these corpora we explored the system they followed to make their corpora. In addition, we have also taken a look at the data distribution of these corpora. And we have used the state-of-the-art Statistical Machine Translation system to evaluate their performance on machine translation.

We outline this paper as follows: we explained the main concepts and a short overview in section 2, then we detailed the review process in section 3 and presented the results and discussion of the review in Section 4. We added Appendices in Section 6.

2 Background

Statistical Machine Translation (SMT) needs considerably good corpus having large amounts of text data to produce good translations. By searching through different libraries, websites for Bangla↔English parallel corpora, we merely found thirteen available corpora. Corpus construction phases are given below.

2.1 Sources of data

By searching through different libraries, websites for Bangla English parallel corpora but we merely found thirteen Bangla↔English available parallel corpora. Some corpora follow newspaper as a corpora source, and some follow the website and Wikisource. Overall we noticed maximum data are collected from online.

[4] *en.wikipedia.org/wiki/Bengali_language*

2.2 Data preprocessing

Since the individual sources of parallel texts differ in many aspects, a lot of effort was required to integrate them into a common framework. We couldn't get more details about preprocessing. From table 6, we get an overview of corpora of how they preprocessed themselves.

2.3 Data distribution

To review every individual corpus, we tried to find out several linguistics features such as sentence fragments, total words, unique words, average words per sentence, words ratio, lexical diversity, etc. With this analysis, we were able to observe ins and out overall qualities, quantities, difficulties, characteristics of all available corpora.

2.4 Evaluation metrics

By using Statistical Machine Translation System, we evaluated translation results by three metrics: BLEU (Papineni et al., 2002), NIST (Przybocki et al., 2009), TER (Snover et al., 2006). BLEU scores were calculated using *multi-bleu.perl* in the Moses toolkit (Koehn et al., 2007). NIST scores were calculated using *mteval-v13a.pl* in the Moses toolkit. To get the proper scores of NIST, you must have to wrap translated file into sgm format by using *wrap-xml.perl*. TER (Translated Edit Rate) measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation. TER scores were calculated using *tercom.version.jar* [5] which use JDK and JRE to run.

In this paper, we reviewed all available corpora of Bangla ↔ English we had found. The following section is devoted to data sources, distribution, evaluation. Thereafter, we discuss provided tools and, finally, we present our plans for future work.

3 Research

In this section, we will describe: Purpose of the Review, Searching parallel corpus, Finding out available relevant corpora, Data Sources and Distribution, Quality Appraisal and Data Evaluation.

We created seven research questions, showed in *Table 1*, to guide this Review as an attempt to find out

[5] http://www.cs.umd.edu/s̃nover/tercom/

Table 1 Research Questions.

ID	Questions
RQ1	How many Bangla↔English parallel corpora were found?
RQ2	Are all the founded corpora publicly available?
RQ3	What are the sources of each available parallel corpus?
RQ4	How data are preprocessed to build corpus?
RQ5	How data are distributed in all available parallel corpora?
RQ6	How each corpus is evaluated via Statistical Machine Translation System?
RQ7	Analyze the statistics of both linguistics features and SMT of available corpora.

all available Bangla ↔ English Parallel corpora, tools, algorithms and systems work.

3.1 Data Retrieval

In 1991, the first attempt was made to develop *Anglabharti* (Sinha et al., 1995), an Bangla↔English Machine Translation system as part of the English-to-Indian Languages system. After that, the number of corpora had increased slowly and reached at fourteen. Book chapters were also excluded from the search. The searches were conducted in five didgital Libraries, ACM Digital Library[6], IEEE Explore[7], Science Direct – Elsevier[8], Springer Link[9] and ACL[10]. To execute this task, a conceptual research string was developed containing the main keyword of the theme. We executed the search strings on February 18, 2020 in each digital library and the results are presented in *Table 2*.

In ACM library's search engine, we searched all of the strings mentioned in *Table 2* by filtering 'Machine Translation' domain in common. If we search without having any domain, engine will show different results and the number of irrelevant papers will be huge because of their searching algorithm.

In SpringerLink, we refined the search contents from given options on the search page and counted only conference papers.

Both ACM and SpringerLink's search engine searches by words in contrast to full strings which is a major problem for getting weird results. For this reason, a huge number of papers come up on the search results, where a maximum of them are not relevant according to the full strings.

[6] https://dl.acm.org
[7] https://ieeexplore.ieee.org/Xplore/home.jsp
[8] https://www.elsevier.com
[9] https://link.springer.com
[10] https://www.aclweb.org/anthology/

Elsevier's searching algorithms seems messy to us because it showed the same results, again and again, despite changing search strings.

3.2 Finding out available relevant corpora and their papers

We applied inclusion and exclusion criteria which is described in *Table 3*, to be explicit about the corpora we considered in our review. We found fourteen relevant corpora. Only nine of them had papers.

4 Results and discussion

This section answers and discusses all the research questions one by one from *Table 1* with a detailed analysis.

4.1 RQ1 - *How many Bangla↔English parallel corpora were found?*

We have found fourteen Bangla↔English Parallel corpora by searching through different libraries. Nine of them are found on Opus[11] named *GlobalVoices*, *Gnome*, *JW300*, *KDE4*, *OpenSubtitles*, *QED*, *Tanzil*, *Tatoeba*, *Ubuntu*. Opus tries to improve the situation by compiling additional data sets on a large scale in order to provide data for many other, often under-resourced languages and domains. The overall goal of the Opus project is to make parallel resources freely available, especially emphasizing the support of low density languages which is really helpful for the researcher (Tiedemann, 2012). Others are given below.

SUPara corpus (Al Mumin et al., 2012) is distributed through the Computer Science and Engineering (CSE) department of Shahjalal University of Science and Technology (SUST).

EMILLE corpus (Baker et al., 2002) developed through the "Enabling Minority Language Engineering" project, which was undertaken by the universities of Lancaster and Sheffield.

ILMPC (Indic Language Multilingual Parallel Corpus) is introduced on Workshop on Asian Translation (WAT 2018) (Nakazawa et al. (2019), Banerjee et al. (2018)).

Pan Treebank Bangla-English parallel corpus (**?**, **?**) (PTB) is developed by PAN Localization Project[12].

SIPC (Six Indian Parallel Corpora) is constructed via CrowdSoucring by using Amazon's Mechanical Turk (MTurk) (Post et al., 2012).

[11] opus.nlpl.eu
[12] https://www.panl10n.net/

Table 2 Number of papers retrieved in each Digital Library after search strings execution.

	Digital Library					
Term Search	ACM	IEEE	Springer	Elsevier	ACL	Total
Machine Translation	1,129	4201	76,474	283	387	82474
Machine Translation of Low Resources	738	60	18,079	2,656	9	21542
Machine Translation of Bangla	6	32	141	2,656	0	2835
Machine Translation of Bengali	7	21	213	2,656	0	2897
Statistical Machine Translation of parallel corpus	1129	121	1,703	2,656	2	5611
Neural Machine Translation of parallel corpus	1129	30	708	2,656	2	4525
SMT of Bangla parallel corpus	990	1	10	2,656	0	3657
NMT of Bangla parallel corpus	978	0	1	2,656	0	3635
Review paper of Low Resources	1,056	313	57,213	2,669	1	61252
Review paper of Bangla MT	1,052	0	6	2,669	0	3727
Review paper of Parallel Corpora MT	1,085	0	236	2,670	1	3992

Table 3 Inclusion and Exclusion Criteria defined for finding out relevant corpus and paper.

Inclusion Criteria	Exclusion Criteria
Papers written in English	Papers written in other languages rather than English
Available Bangla ↔ English parallel corpus	All others parallel corpus without Bangla or Bengali
Bangla↔English bilingual corpus	All kinds of monolingual corpus

Table 4 List of websites where corpora were available for download (Sorted Alphabetically)

Corpus Name	**Available On**
EMILLE	Personally Collected
GlobalVoices	Opus/global-voices Casmacat/global-voices
Gnome	Opus/gnome
ILMPC	WAT/indic-multilingual
JW300	Opus/jw300
KDE4	Opus/kde4
OpenSubtitles	Opus/open-subtitles Opus/open-subtitles-alt
QED	Opus/qed QCRI
SUPara	supara-github
SIPC	sipc-github
Tanzil	Opus/tanzil TanzilNet
Tatoeba	Opus/tatoeba
Ubuntu	Opus/ubuntu

4.2 RQ2 - *Are all the founded corpora publicly available?*

All of the founded corpora except *Pan Treebank*, are available free of charge for educational and research purposes, however, the license allows collecting statistical data and making short citations.

Maximum corpora are found as *tmx*, *moses* and *txt* formats in different websites. Two of them are not available on internet. So we collected them personally from respected Authors. Some of the corpora have software that uses their own corpus to translate from one language to another. All the information are provided in the *Table 5*. All datasets' links are provided in Appendix at the last of the paper.

4.3 RQ3 - *What are the sources of each available parallel corpus?*

Here is a discussion about the source of the corpora that we've worked with. All source link is appended in Appendix B.

In *JW300* corpus (Agić and Vulić, 2019), data collected from online website *jw.org*. The vast majority of texts collected from the magazines *Awake!* and *Watchtower*. In this corpus, multilingual articles are mainly translated from the bible to 300 languages.

In *SIPC* corpus (Post et al., 2012), they apply an established protocol for using *Amazon's Mechanical Turk* (MTurk) to collect parallel data to train and evaluate translation systems for six Indian languages. They investigate the relative performance of syntactic translation models and explore the impact of training data quality on the quality of the resulting model. Finally, they release the corpora to the research community under the Creative Commons Attribution-Sharealike 3.0 Unported License.

QED or *AMARA* (Abdelali et al., 2014) corpus is an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated over the AMARA web-based platform. Here maximum contents are collected from *Khan Academy*, *Coursera*, *Udacity*, *TED Talks* and *ELRA* website.

In *OpenSubtitles* corpus (Lison and Tiedemann, 2016), the dataset consists of a database dump of the *Open-*

Table 5 Sources of all available parallel corpora

	Corpora										
Sources	GlobalVoices	GNOME	ILMPC	JW300	KDE4	OpebSubtitles	QED	Supara	SIPC	Tatoeba	Tanzil	Ubuntu
Amazon's Mechanical Turk									✓			
Budget speech								✓				
Commercial policy								✓				
Essential interfaces								✓				
European Central Bank(ECB)		✓		✓	✓						✓	✓
Educational video subtitles							✓					
Education policy								✓				
Government official website								✓				
Holy Book				✓							✓	
Khan Academy							✓					
Movie subtitles						✓						
Magazines				✓			✓					
Novel and Feature								✓				
Newspaper	✓	✓		✓	✓			✓		✓	✓	✓
OpenSubtitles						✓						
Online sources.	✓	✓	✓				✓			✓		✓
Online educational content								✓				
Prime Minister's speeches								✓				
Rules and Procedure of Parliament								✓				
TED							✓					
Udacity							✓					
WikiSource	✓	✓		✓	✓					✓	✓	✓
Website	✓	✓		✓		✓	✓		✓	✓	✓	✓
Websites of several companies								✓				
Wikipedia and Banglapedia								✓	✓			

Subtitles.org repository of subtitles, comprising a total of 3.36 million subtitle files covering more than 60 languages. This is a slightly cleaner version of the subtitle collection using improved sentence alignment and better language checking.

SUPara (Al Mumin et al., 2012) have used texts from some sources that are either publicly available or granted permission from respective copyright holders. They collect data from novel and feature. Some Documents containing Prime Minister's speeches, budget speech, commercial policy, etc are obtained from the government official website. Several essayistic texts are collected from different online newspapers, e.g., Bangladesh Sangbad Sangstha(BSS), BDNews24.com. Many other documents are obtained from websites of several companies like Grameen Phone, Bangladesh Parjatan Corporation and so on. They also generated some documents by mining data from *Wikipedia* and *Banglapedia*.

Tanzil (Tiedemann, 2012) is a collection of Quran translations, and all data are collected from the Quran and their own *website*.

GlobalVoices (Tiedemann, 2012) collected data mainly from *Global Voice* website. *Tatoeba* (Tiedemann, 2012) collected data from *Tatoeba* website.

KDE4, *Ubuntu*, *GNOME* corpora are originally mentioned in OPUS's *website*. We didn't get much idea about its source from (Tiedemann, 2012) paper.

For all available corpora from Opus, website also provides pre-compiled word alignments and phrase tables, bilingual dictionaries, frequency counts, and these files are found there.

4.4 RQ4 - *How data are preprocessed to build corpus?*

We couldn't find more details on preprocessing on the corpora. In the following table-6, we tried to give a rough idea of the few preprocessing terms corpora that follow. Here, the following steps have been applied as preprocessing on the English and Bangla documents for the available corpora:

Cleaning up Documents - This cleaning up means that the various formats, for example, rtf, doc, and pdf, are converted to plain text files. Tagged files like HTML and PHP files are normalized by deleting tags and then converted to plain text files.

Encoding and Markup - The texts are encoded according to international standards by using UTF8 (Unicode). For Bangla documents, they have used the 'Nikosh' converter to encode all formats into Unicode.

Alignment - The alignment of translated segments with source segments is essential for building parallel corpora. Each file in a sub-directory is aligned separately with its translation to keep alignment errors at a low level.

Tools - Some corpora used various tools for preprocessing the data. In particular, they applied various types of open-source software and free research tools. These tools include sentence splitter, word histogram generator, Unicode converter, etc.

Subtitle conversion - In table 6, we can see only the OpenSubtitles corpus follow these preprocessing terms. OpenSubtitles does not enforce any particular encoding format on the subtitles uploaded by its users. The most likely encoding for the file must therefore be determined based on various heuristics. When several alternative encodings are admissible, the 'chardet' library is applied to determine the most likely encoding given the file content (Li and Momoi, 2001).

Sentence segmentation and tokenization - In table 6, we can see most of the corpora follow these preprocessing terms except the Supara corpus.

Correction of OCR and spelling errors - Many subtitles in our dataset are automatically extracted via Optical Character Recognition (OCR) from video streams, leading to a number of OCR errors.In table 6, we can see most of the corpora follow these preprocessing terms except the Supara corpus, the QED corpus, the ILMPC corpus.

Inclusion of meta-data - This preprocessing step is to generate the meta-data associated with each subtitle.

4.5 RQ5 - *How data are distributed in all available parallel corpora?*

4.5.1 Coverage of language

JW300 is the largest having huge coverage of languages, almost 380 languages mentioned in *Table 7*. Tatoeba covers the second most languages. But Bangla $\leftrightarrow$ English parallel corpus is tiny having 5,120 sentences mentioned in *Table 8*. Though GlobalVoices has the lowest language coverage among other corpora of Opus, Bangla $\leftrightarrow$ English side is pretty huge having 137,620 sentences. The number of language coverages of other Opus corpora is given in *Table 7*. All available corpora of Opus are bidirectional. Which means, Bangla to English or English to Bangla translation both are possible.

ILMPC corpus covers eight languages. These are Bangla, Hindi, Malayalam, Tamil, Telegu, Sinhalese, Urdu and English. This corpus is used for the pilot as well as multilingual English-Indic or Indic-English Languages sub-tasks. It is a collection of 7 bilingual parallel corpora of varying sizes, one for each Indic language and

Table 6 Statistics overview of Corpora of how they preprocessed themselves.

	Corpus Name				
Preprocessing Terms	GlobalVoices GNOME JW300 Tanzil Ubuntu	OpebSubtitles	QED	Supara	ILMPC
Cleaning up Documents				✓	
Encoding and Markup				✓	
Alignment				✓	
Tools				✓	
Subtitle conversion		✓			
Sentence segmentation	✓	✓	✓		✓
Tokenization	✓	✓	✓		✓
Correction of OCR	✓	✓			
Spelling errors	✓	✓			
Inclusion of meta-data		✓			
Training			✓		✓

Table 7 *Bird's-eye-view of founded Parallel Corpora (Sorted Alphabetically)*

Corpus Name	Languages	Files	Sentence Fragments	Tokens
Emille	14			67M
GlobalVoices	41	224,096	4.93M	88.71M
Gnome	187	113,344	58.12M	267.27M
ILMPC	8	3,200		4.64M
JW300	**380**	1,285,939	105.11M	1.95G
KDE4	92	75,535	8.89M	60.75M
OpenSubtitles	62	**3,735,070**	**3.35G**	**22.10G**
QED	225	271,558	30.93M	371.76M
SUPara	2	80	0.02M	0.45M
SIPC	6		0.15M	1.6M
Tanzil	42	105	1.01M	22.33M
Tatoeba	309	309	7.82M	57.55M
Ubuntu	244	30,959	7.73M	29.84M

English. The parallel corpora are also accompanied by monolingual corpora from the same domain.

SUPara corpus is an Bangla↔English parallel corpus consisting of more than 0.45M words in either languages, which is the largest among freely released corpus of its kind.

EMILLE corpus consists of a series of monolingual corpora for fourteen South Asian Languages and a parallel corpus of English and five of these languages.

SIPC is a collection of parallel corpora between English and six languages from the Indian subcontinent: Bangla, Hindi, Malayalam, Tamil, Telugu, and Urdu.

4.5.2 Arrangement of files

OpenSubtitles covers a total of 152,939 movies or TV episodes (as determined by IMDb identifier). 70% of the IMDb identifiers are associated with subtitles in at least two languages, 44% with at least 5 languages, 28% with at least 10 languages, and 8% with at least 20 languages. However, the reason behind having huge files of OpenSubtitles is, for each movie or TV episode may have more than one file in case of multiple CDs.

QED or AMARA corpus covers 44,620 online educational videos, TV shows such as Khan Academy, TED Talks, Udacity, Coursera and generates around 271,558 files from their subtitle files. JW300 generates files from Bibel, New Testament, Awake! and Watchtower magazines by segmenting each chapter or article into a reasonable size. Tanzil builds 80 files by splitting the Holy Quran into 80 sections. SUPara corpus contains altogether 80 document pairs that consist of literature, journalistic, instructive, administrative, external communication. Other corpora mentioned in *Table 7* follow the same rule to generate files.

4.5.3 Count of sentences or fragments

A sentence fragment is a word, phrase, or dependent clause that is punctuated as a sentence, but the subject, verb, or both may be missing. Though sentence fragments may be used for effect in certain types of writing, fragments are generally not used in academic or pro-

Table 8 *Statistics Analysis of Available English ↔ Bangla Parallel Corpora (Sorted Alphabetically)*

	Sentences	Words		Unique Words		Average Words per Sentence	
Corpus Name	En or Bn	En	Bn	En	Bn	En	Bn
Emille	6,375	89,027	90,062	6,636	10,816	13.965	14.127
Global Voices	137,620	2,536,451	2,269,045	**88,602**	**130,606**	18.4308	16.4877
Gnome	132,481	637,363	619,182	6,869	7,147	4.811	4.674
ILMPC	337,428	2,260,636	1,840,722	42,510	83,418	6.710	5.455
$JW300_1$	366,972	5,180,241	5,074,551	43,781	80,447	14.116	13.828
$JW300_2$	370,948	**5,890,630**	**5,083,847**	57,461	80,575	2.332	13.705
KDE4	36,381	149,273	129,159	14,174	10,869	4.103	3.550
OpenSubtitles	**413,602**	2,401,653	1,974,181	42,432	84,343	5.807	4.773
QED	2,043	71,695	245,548	8,167	26,674	**35.093**	**120.190**
SUPara	21,158	244,539	202,866	14,571	22,456	11.56	9.59
SIPC	20,788	290,972	240,077	16,594	13,501	13.997	11.549
Tanzil	187,052	4,391,125	4,185,894	17,600	17,925	23.475	22.378
Tatoeba	5,120	24,656	22,321	2,082	3,345	4.816	4.360
Ubuntu	5,634	21,791	17,900	5,619	4,053	3.868	3.177

fessional writing. Some of the corpora especially Opus used fragments of sentences. Others (such as ILMPC, SUPara) used full sentences, instead of using a fragment.

We found two different corpora of JW300, shown in *Table 8*. Every corpus except $JW300_1$ and $JW300_2$[1] have same number of sentences or fragments of sentence. These two corpora are not even properly aligned with each other which is so much important for translation. For example, Bangla document's fifth line is not aligned with the English document's fifth. There exists excessive new lines (empty lines) instead of proper alignment.

OpenSubtitles has the maximum number of sentence fragments around 3.35G because of the availability of online subtitles' sources. Also, Bangla ↔ English side of OpenSubtitles has the largest number of sentences. Overall count of sentence fragments for EMILLE and ILMPC are not mentioned in the dataset's description. SUPara has only 0.02M sentences because it covers only two languages. But SUPara is one of the finest datasets of it's kind. Count of sentences or fragments of other corpora is mentioned in *Table 8*.

4.5.4 Smallest unit

Token is the smallest unit that each corpus divides to. Typically each word form and punctuation (comma, dot, . . .) is a separate token. Therefore, corpora contain more tokens than words. Spaces between words are not tokens. But in *Table 8*, the term "Token" actually refers to the total number of words in the corpus, of course excluding punctuations. OpenSubtitles has the largest number of tokens nearly 22.1G because of its data collection policies. SUPara has 0.45Million tokens because it covers only two languages. Others are mentioned in *Table 7*.

"Words" which is mentioned in *Table 8*, denotes the total number of barely Bangla words in Bangla side and English words in English side corpus individually. Some corpora (such as Gnome, Ubuntu, QED) have Bangla sentences having a lot of English words. We excluded English words from them and counted only Bangla words. $JW300_2$ has the highest number of words in both English and Bangla side. QED or *AMARA* corpus (Abdelali et al., 2014) has 71,695 English words in contrast to 245,548 Bangla words. This means, almost every sentence of Bangla corpus has more words than the aligned English sentence. Translated English sentence is incomplete for almost every sentences of Bangla. That's why QED has the highest "average words per sentence" in Bangla side which is 120.190 whereas aligned English side has an average 35.093 words per sentence. Number of "Unique Words" of each corpus are given in *Table 8*. First, we tokenized them into merely Bangla words for Bangla side and English words for English side and then used set to find out unique words. GlobalVoices has the highest unique words for both in Bangla and English side.

4.5.5 Diversities

Lexical diversity is one aspect of "lexical richness" and refers to the ratio of different unique word stems (types) to the total number of words. It refers to "the range of different words used in a text, with a greater range indicating a higher diversity". Details of lexical diversity of available corpora are shown in *Fig 1*. Big score of *Tanzil* refers that on average each vocabu-

[1] $JW300_2$ has 370,948 English and 2,525,512 Bangla Lines. It isn't properly aligned with sentence by sentence.

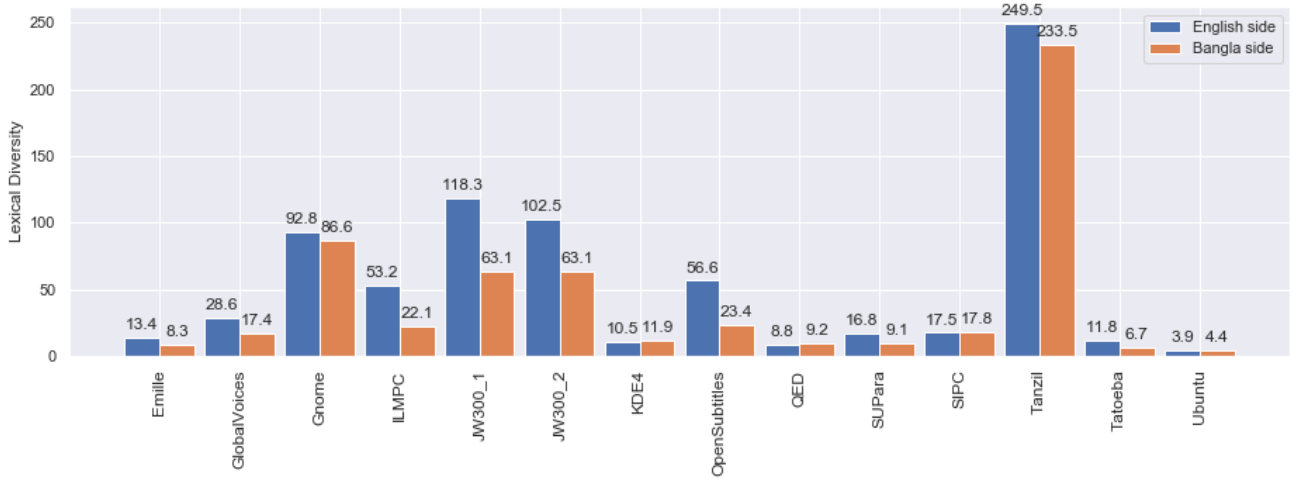

Fig. 1 Lexical Diversity of available corpora

lary item appears very frequently for both English and Bangla sides. Ubuntu got the lowest score which means items appear very rarely on sentences for both English and Bangla side.

4.6 RQ6 - *How each corpus is evaluated via Statistical Machine Translation System?*

In this paper, we have gone through two major Machine Translation paradigms: Phrase-based Statistical Machine Translation (SMT) and recent advent– Neural based Machine Translation (NMT). In MT research, SMT (Koehn and Knight, 2003) has been considered as the state-of-the-art technology until the advent of NMT (Kalchbrenner and Blunsom (2013), Sutskever et al. (2014), Cho et al. (2014)) recently, which shows an improved result for many high-resource language pairs (Sennrich et al., 2016). However, for low-resource settings, SMT is still considered as state-of-the-art technology since NMT failed to show improved performance in this scenario (Koehn and Knowles (2017), Östling and Tiedemann (2017), Mumin et al. (2019)).

We followed Mumin et al. (2019) for the preprocessing, training, tuning and system configuration.

Preprocessing - We normalized both English and Bangla sides using standard *normalize-punctuation.perl* of the Moses toolkit (Koehn et al., 2007). We further tokenized English side with *tokenizer.perl* [13] in the Moses-decoder and Bangla side by using our own tokenizer[14]. Then we used *clean-corpus-n.perl* to create clean corpus of both Bangla and English.

Training - We built word alignments using *train-model.perl* and symmetrized them using grow-diag-final-and heuristic (Koehn et al., 2007). We extracted phrase pairs up to a maximum length of eight words. We scored these phrase pairs using maximum likelihood, thus obtaining a phrase table where each phrase-pair has the standard translation model features. We also built a lexicalized reordering model by using *msd-bidirectional-fe*.

For language modeling, we trained a separate 5-gram LM model for both English and Bangla sides of the training bitext using *lmplz* to *arpa* format and then using *build_binary* to *blm* format. For Bangla, we used SUMono(Al Mumin et al., 2014), a monolingual corpus. And for English, we used Europarl(Koehn, 2005). Finally, we built a large joint log-linear model, which used standard SMT feature functions: language model probability, word penalty, the parameters from the phrase table, and those from the reordering model.

Tuning - We tuned the weights in the log-linear model by using SUparadev dataset. We did not have much time to use ilmpc-dev or sipc-dev (*Table 10*) for tuning purpose. These tasks will be accomplished in further research.

We used the phrase-based SMT model as implemented in the Moses toolkit (Koehn et al., 2007) for translation, and reported evaluation results over three test sets, suparatest in *Table 10*, ilmpc-testest in *Table 11* and sipctestset in *Table 12*.

SUPara corpus got the best result while using suparatestset and sipctestset for the statistical machine translation system. But with suparatestset, deviation of BLEU scores among corpora is really high in both En ↔ Bn and Bn ↔ En side. Some corpora got 0.00 BLEU scores because of their unorganized dataset, ratio of Bangla and English words per sentence, high lexi-

[13] https://github.com/moses-smt/mosesdecoder /tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl
[14] github.com/masumahmedeesha/reviewOfCorpora

Table 9 Statistics of available development and test sets among founded corpora

Corpus Name	Data sets	Sentences	Tokens(EN)	Tokens(BN)
ILMPC	Development	500	8,050	6,786
	Test	1,000	14,755	12,548
SUPara	Development	500	9,865	7,902
	Test	500	9,926	7,938
SIPC	Development	914	13,363	10,545
	Test	1,001	15,134	11,756

Table 10 Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **SUParatestset** for evaluating]

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	2.44	2.13	101.96	2.97	2.82	87.25
Gnome	0.92	1.64	107.43	1.65	2.18	88.95
GlobalVoices	9.80	4.21	79.21	10.58	4.73	75.51
ILMPC	4.79	3.06	88.47	7.70	3.89	81.27
KDE4	1.34	1.63	109.10	1.56	2.21	90.97
OpenSubtitles	5.15	3.052	87.51	7.83	1.00	80.89
QED	0.00	1.12	105.16	0.70	1.59	93.81
SUPara	**15.24**	**5.06**	**72.47**	**16.85**	**5.70**	**68.51**
SIPC	3.29	2.62	90.13	5.10	3.22	84.71
Tatoeba	0.74	1.21	108.74	1.19	0.87	89.00
Tanzil	1.35	1.64	96.11	2.48	2.20	93.26
Ubuntu	1.04	1.39	108.95	1.30	1.89	92.08

Table 11 Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **ILMPCtestset** for evaluating]

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	1.58	1.81	101.25	2.64	2.11	101.79
Gnome	2.70	1.41	101.14	1.71	1.79	97.90
GlobalVoices	3.94	2.87	84.02	7.01	3.53	85.34
ILMPC	6.40	3.23	80.52	11.92	4.33	79.59
KDE4	2.82	1.40	103.20	1.71	1.67	110.42
OpenSubtitles	**19.16**	**4.69**	**65.07**	**33.18**	**6.66**	**58.67**
QED	0.71	1.39	94.84	2.23	1.92	106.53
SIPC	1.27	1.71	95.08	2.74	2.25	93.63
SUPara	4.18	2.93	85.01	8.04	3.78	83.27
Tatoeba	1.75	1.90	95.12	3.34	2.30	90.86
Tanzil	1.56	1.63	92.65	2.51	2.15	103.68
Ubuntu	1.72	1.17	102.57	1.29	1.52	102.00

cal diversity, etc. By using ilmpctestset, OpenSubtitles got the best result. But still got the huge deviation problem in BLEU scores section.

4.7 RQ7 - *Analyze the statistics of both linguistics features and SMT of available corpora*

OpenSubtitles has the maximum number of sentences or fragments of sentences mentioned in *Table 8*. But not all sentences are full, there exist fragments of sentences that affect Bangla English words ratio. $JW300_2$ has the largest number of "Words" in both English and Bangla sides although this corpus is not properly aligned with Bangla and English side's sentences. *GlobalVoices* has the largest number of "Unique Words" in both English and Bangla sides. Although *QED* has the maximum number of "Average Words per sentence" in both English and Bangla sides, Bangla↔English isn't fully translated. Bangla→English words ratio is 3:1. Bangla

Table 12 Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **SIPCtestset** for evaluating]

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	0.37	1.14	113.12	1.26	1.70	93.35
Gnome	0.77	1.15	113.56	0.82	1.48	93.32
GlobalVoices	7.35	3.35	86.04	8.66	4.13	82.40
ILMPC	2.90	2.24	98.49	4.50	3.08	88.31
KDE4	0.67	1.08	117.09	1.31	1.55	94.46
OpenSubtitles	2.54	2.15	98.78	4.44	3.11	88.20
QED	0.00	0.68	116.09	0.36	1.12	95.98
SIPC	7.36	3.29	85.60	7.02	3.56	87.07
SUPara	**8.87**	**3.54**	**85.24**	**9.88**	**4.49**	**79.45**
Tatoeba	0.28	0.68	118.39	0.38	0.81	95.43
Tanzil	0.00	0.84	109.87	1.16	1.33	96.23
Ubuntu	0.00	0.84	119.05	0.68	1.17	94.61

sentence has more words and too long whereas English sentence is very short and doesn't have a full translation of that Bangla sentence. Ubuntu has the best lexical diversity score but it is the smallest among available corpora.

We used SUParadev for tuning purposes. By using SUParatestset for evaluation of three metrics mentioned in Table, Statistical Machine Translation system generates the best results for SUPara in both En to Bn and Bn to En sides. With ilmpctestset, OpenSubtitles achieved the best results mentioned in Table as well as by using sipctestset, SMT again generates the best performance for SUPara.

5 Conclusion

We summarized in this study about the current state of existing available Bangla ↔ English parallel corpora. This review found fourteen corpora. Except for Pan Treebank, other thirteen corpora are publicly available free of charge for educational and research purposes, however, the license allows collecting statistical data and making short citations. Available corpora were analyzed by finding common features listed them into tables and also into some figures to visualize the quality of each corpus.

Besides the paradigms and their approaches, we were able to enlighten our results with three evaluation metrics by following phrased based statistical machine translation methods. This helped our way to dig deeper into the corpus and find better one. We used three different test sets to make sure the correctness of our SMT model.

For future research, we would like to train all corpora with ILMPCdev and SIPCdev, and then evaluate BLEU, NIST and TER. Also we want to use Neural Machine Translation methods for far better results to judge the quality of corpora. We would like to add "Word embedding" for each corpus which uses continuous-space representation for natural language.

6 APPENDICES

6.1 Appendix A

Table 13 List of all data sets' links are given.

6.2 Appendix B

Table 14 List of all sources' links are given.

Acknowledgement

The first author is grateful to Information and Communication Technology (ICT) Division, Government of People's Republic of Bangladesh for the grant to do this research work.

Funding Information

Mohammad Abdullah Al Mumin's work has been supported by ICT Division, Ministry of Posts, Telecommunications and IT, Government of the People's Republic of Bangladesh [Order No: 56.00.0000.028.33.077.17-78, date: 02.04.2018].

Table 13 List of dataset links to download available parallel corpora

Corpus Name	Dataset link
Opus/global-voices	http://opus.nlpl.eu/GlobalVoices.php
Opus/gnome	http://opus.nlpl.eu/GNOME.php
gnome-org	https://l10n.gnome.org/
WAT/indic-multilingual	http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html
Opus/jw300	http://opus.nlpl.eu/JW300.php
Opus/kde4	http://opus.nlpl.eu/KDE4.php
Opus/open-subtitles	http://opus.nlpl.eu/OpenSubtitles.php
Opus/open-subtitles-alt	http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php
Opus/qed	http://opus.nlpl.eu/QED.php
QCRI	http://alt.qcri.org/resources/qedcorpus/
supara-github	https://github.com/maamumin/SUPara
sipc-github	https://github.com/joshua-decoder/indian-parallel-corpora
Symfony	https://symfony.com/legacy
Opus/tanzil	http://opus.nlpl.eu/Tanzil.php
TanzilNet	http://tanzil.net/trans/
TranslationTanzil	http://tanzil.net/#19:1
Opus/tatoeba	http://opus.nlpl.eu/Tatoeba.php
Opus/ubuntu	http://opus.nlpl.eu/Ubuntu.php
Translation-LaunchPad	https://translations.launchpad.net/

Table 14 Source link

Source Name	Website link
Amazon's Mechanical Turk	https://www.mturk.com/
Wikipedia	https://www.wikipedia.org/
Khan Academy	https://www.khanacademy.org
Coursera	https://www.coursera.org
Udacity	https://www.udacity.com
TED Talks	http://www.ted.com
ELRA	http://www.elra.org
Banglapedia	https://www.banglapedia.org/
Awake and Watchtower	https://www.jw.org/en/library/magazines/

References

Abdelali A, Guzman F, Sajjad H, Vogel S (2014) The amara corpus: Building parallel language resources for the educational domain. In: LREC, vol 14, pp 1044–1054

Agić Ž, Vulić I (2019) JW300: A wide-coverage parallel corpus for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 3204–3210, DOI 10.18653/v1/P19-1310, URL https://www.aclweb.org/anthology/P19-1310

Al Mumin MA, Shoeb AAM, Selim MR, Iqbal MZ (2012) Supara: a balanced english-bengali parallel corpus. SUST Journal of Science and Technology pp 46–51

Al Mumin MA, Shoeb AAM, Selim MR, Iqbal MZ (2014) Sumono: A representative modern bengali corpus. SUST Journal of Science and Technology 21:78–86

Baker P, Hardie A, McEnery T, Cunningham H, Gaizauskas RJ (2002) Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation. In: LREC

Banerjee T, Kunchukuttan A, Bhattacharyya P (2018) Multilingual indian language translation system at wat 2018: Many-to-one phrase-based smt. In: Proceedings of the 5th Workshop on Asian Translation (WAT2018)

Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259

Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp 1700–1709

Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: MT summit, Citeseer, vol 5, pp 79–86

Koehn P, Knight K (2003) Feature-rich statistical translation of noun phrases. In: proceedings of the

41st Annual Meeting of the association for Computational Linguistics, pp 311–318

Koehn P, Knowles R (2017) Six challenges for neural machine translation. arXiv preprint arXiv:170603872

Koehn P, Federico M, Shen W, Bertoldi N, Bojar O, Callison-Burch C, Cowan B, Dyer C, Hoang H, Zens R, et al. (2007) Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In: Final Report of the Johns Hopkins 2006 Summer Workshop

Li S, Momoi K (2001) A composite approach to language/encoding detection. In: Proc. 19th International Unicode Conference, pp 1–14

Lison P, Tiedemann J (2016) Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles

Mumin M, Seddiqui M, Iqbal M, Islam MJ (2019) shu−torjoma: An english↔bangla statistical machine translation system. Journal of Computer Science 15:1022–1039, DOI 10.3844/jcssp.2019.1022.1039

Nakazawa T, Doi N, Higashiyama S, Ding C, Dabre R, Mino H, Goto I, Pa WP, Kunchukuttan A, Parida S, et al. (2019) Overview of the 6th workshop on asian translation. In: Proceedings of the 6th Workshop on Asian Translation, pp 1–35

Östling R, Tiedemann J (2017) Neural machine translation for low-resource languages. arXiv preprint arXiv:170805729

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 311–318

Post M, Callison-Burch C, Osborne M (2012) Constructing parallel corpora for six indian languages via crowdsourcing. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp 401–409

Przybocki M, Peterson K, Bronsart S, Sanders G (2009) The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. Machine Translation 23(2-3):71–103

Sennrich R, Haddow B, Birch A (2016) Edinburgh neural machine translation systems for wmt 16. arXiv preprint arXiv:160602891

Sinha R, Sivaraman K, Agrawal A, Jain R, Srivastava R, Jain A (1995) Anglabharti: a multilingual machine aided translation project on translation from english to indian languages. In: 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, IEEE, vol 2, pp 1609–1614

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas, vol 200

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112

Tiedemann J (2012) Parallel data, tools and interfaces in opus. In: Chair) NCC, Choukri K, Declerck T, Dogan MU, Maegaard B, Mariani J, Odijk J, Piperidis S (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey