Developing the Bangladeshi National Corpus–a Balanced and Representative Bangla Corpus

Conference Paper · December 2019

DOI: 10.1109/STI47673.2019.9068005

4 authors, including:

Khan Md Anwarus Salam
IBM Japan
19 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Mahfujur Rahman
1 PUBLICATION **0** CITATIONS
SEE PROFILE

2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 24-25 December, Dhaka

Developing the Bangladeshi National Corpus- a Balanced and Representative Bangla Corpus

Khan Md Anwarus Salam*, Mahfujur Rahman°, Md Mahfuzus Salam Khan‡
Chief Technology Officer*, Research Coordinator°, Chief Executive Officer‡
Dream Door Soft Ltd.
Dhaka, Bangladesh
{anwar*, risad°, mahfuz‡}@dreamdoorsoft.com

Abstract— **The need for a balanced, representative national scale corpus has been skyrocketing for the already 'low resource' tagged language-Bangla. Many sporadic empirical works have been done so far in the field of NLP and Computational Linguistics yet, and these are never enough. Moreover, none of these works can bear the best fruit without the help of a standard corpus. To address these issues, the goal of this research work was set to compile the Bangladeshi National Corpus (BDNC). This paper proposes the development process of the BDNC (first phase- Bangla monolingual corpus). In this work, the whole task was divided into three major phases, where the goal of the first phase is to build a representative monolingual corpus that will include at least 100 million Bangla words. Whereas, in the second phase, there will be a sub-corpora that will consist of a parallel corpus having 1 million words in Bangla and English. However, at the third and final phase, the parallel corpus will incorporate 15 foreign languages (including English) comprising a weighted corpus size of at least 15 million words.**

Keywords— Bangla, Corpus, balanced, representative, monolingual corpus, multi-lingual corpus, translation corpus, parallel corpus.

I. INTRODUCTION

Bangla, also known as Bengali, is the national language in Bangladesh and even a mother tongue in the Indian state of West Bengal. Bangla has more than 260 million speakers worldwide, and it is the sixth most spoken language in the world [22]. However, Bangla is still considered a low-resource language because of the unavailability of a balanced corpus with digitally accessible resources.

Corpus is a much needed structured data set of language instances that work as a heart for many tools of Natural Language Processing (NLP). Researchers of different scientific domains also find it as a useful tool. However, for many practical reasons, despite having such advantages, there are not many instances of the balanced, representative corpus being developed for Bangla. Bangla is already considered a low resource language as far as language technology concerns and the lack of having a standardized corpus is also a reason behind it. Needless to say, the relation between these two problems can be labelled as an example of bidirectional causation.

Throughout the document, we'll be discussing our approaches to build the corpus and methods that we'll be using in the course of corpus creation.

Bangladesh is often considered as one of the fastest emerging nations in the world in terms of economic growth. Besides, the country has a reputation in utilizing IT in the most creative and effective ways to solve many of its problems. Yet, the challenges of the 4th Industrial revolution are enormous to countries like Bangladesh. For Bangladesh, the readiness for Industry 4.0 means, being equipped with some sets of prerequisites that include- Bangla Language Processing (NLP) techniques, tools, and various AI solutions (Bangla enabled) among other major phenomena. A well-made, maintained balanced and representative corpus gives a solid ground for Bangla NLP researches and other related fields, thus fostering the backbone development required to face the challenges of Industry 4.0.

There are already some notable works being done in the field of corpus creation. Salam, Yamada and Nishino [2] proposed a balanced corpus for Bangla language for the first time. Sarkar, Pavel and Khan [17] attempted an automatic corpus creation process where they collected all the already available texts from the web and other offline resources as the text source of the corpus. Another attempt was the creation of CIIL corpus Dash and Chaudhuri [16], which was actually a collection of corpus or corpora of nine Indian languages including Bangla. The corpus has a size of 3 million words. Mumin, Shoeb, Selim and Iqbal have built a corpus titled SUPara [18], which was an English-Bangla parallel corpus in 2011. The corpus has more than 200000 words in either language. The same authors created another corpus named SUMono [14] in 2013 which was actually a monolingual corpus consisting of a word size of more than 27 million. This corpus was created following the framework of the American National Corpus. Another such parallel corpus creation attempt was carried out recently though the data collection method was crowd-sourcing [22]. This Bangla-English corpus has a total of 517 Bangla sentences and 2143 corresponding English translations while every Bangla sentence was translated by an average of 4 times via crowd-sourcing. Shamshed and Karim [20] proposed a corpus intended for an efficient way of information retrieval. A newspaper specific corpus was created by Majumder and Arafat [19] where the authors used texts from a Bangla daily newspaper for a particular year. Khan, Ferdousi and Sobhan [15] created another Bangla corpus titled "BDNC01". The size of the corpus was 12 million words and the texts were collected from some of the Bangla daily newspapers and some Bangla literature.

II. DEVELOPMENT PHASES OF BDNC

A. First Phase (Bangla Monolingual Corpus)

A monolingual corpus can be either general or special. In our scope, we are up to build the monolingual corpus as a general one so that it can eventually represent the national

978-1-7281-6099-3/19/$31.00 ©2019 IEEE

variety of colloquial Bangla language [13]. However, the corpus, in the long run, will also reflect the diachronic features of Bangla language. Below is the flow-chart showing the principal steps that we have considered while developing our corpus (mono-lingual) in the first phase.

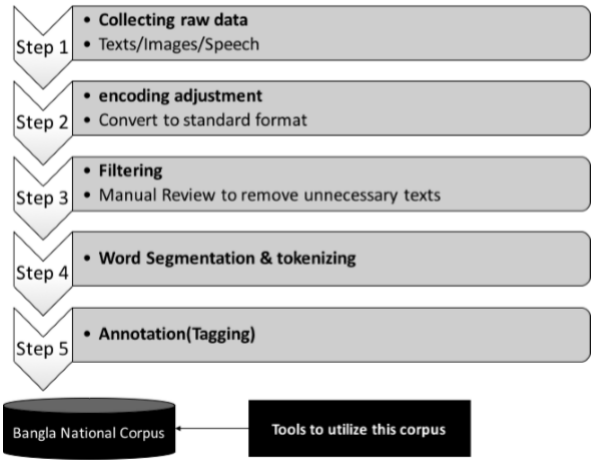

Fig. 1. The development process of Bangla corpus

B. Second & Third Phase (Multilingual Parallel Corpora)

In the second and third phase of the Corpus development task, we will be using the following flowchart as the goal is to develop multilingual parallel corpora.

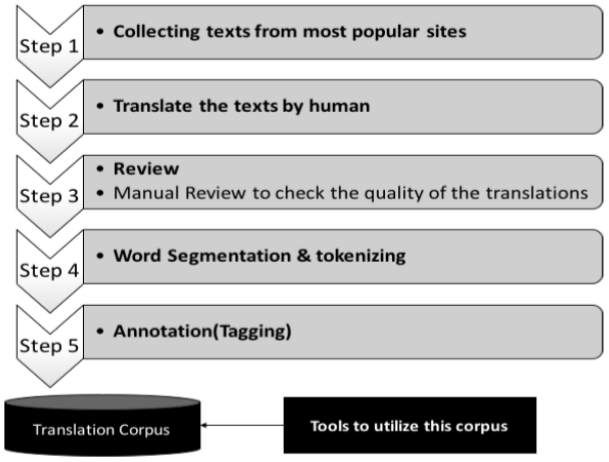

Fig. 2. The development process of parallel corpus

In principle, the goal of the second phase is to translate (human aided) or gather already translated texts (in both Bangla to English and English to Bangla) in order to build a parallel corpus (translation corpus) in Bangla-English.

And the aim of the third phase of the project is to translate manually (human aided) popular website contents and other available resources written in Bangla or English to a number of foreign languages including- Arabic, Bangla, Spanish, French, Mandarin, Japanese, Korean, Hindi, Persian, Burmese, Bhutanese, Urdu, Russian, German, Portuguese and English.

III. CORPUS DESIGN

Before starting with building the actual corpus it is mandatory to design the corpus with proper alignment with the goal and purpose of the corpus itself. We considered two major criteria to design a corpus- one is the Purpose design and another one being Model design. The following table states the different minor notions that we have also considered while designing the criteria above.

TABLE I. CORPUS DESIGN CRITERIA

Purpose Design	Model Design
Scope of usage defining	Corpus Typology design
User defining	Tagset design
Service and QoS design	Storage and Database design

To make a balanced and representative corpus, we are following three independent selection criteria: domain, time and medium [2]. We followed the Chinese SINICA corpus design methodology and added three more attributes, author, writing level and target audience. Table II shows the proposed domain balance percentage.

TABLE II. DOMAIN BALANCE PERCENTAGE

Domain / Source	Percentage
Text Books	20%
Mass Media	20%
Literature	15%
Spoken corpus	10%
Translations	5%

IV. THE DEVELOPMENT PROCESS OF THE BANGLA MONOLINGUAL CORPUS

After designing the purpose and model of the corpus, one can start building the corpus. Following are the steps that we have followed to build the monolingual corpus.

A. Collecting Raw Data

In order to maintain representativeness and to build a balanced corpus, texts are to be collected from various sources that will ensure all the features (both spoken and written forms of colloquial Bangla language used in various domains) and objectives (balanced, having representativeness) that the corpus should hold. The text can be collected in many ways including the followings- using OCR, web-crawling, typewriting, existing electronic text, using STT etc.

- **Using OCR:** Optical Character Recognition (OCR) is considered a way of obtaining electronic texts from books. In this case, human aided proofreading or editing is needed, to correct scanning errors and other technical errors.
- **Typing:** Right now scanner machines and computer programs are not efficient enough at recognizing Bengali texts of different typefaces, lower-quality typography, or handwriting. Therefore, typing can be considered as a solution, though it is a labour-intensive and resource-hungry option. Still, this method is better for leaflets, hand-written items, and recorded speech.

- **Existing electronic texts:** There are many texts already exist in electronic form in Bengali which is a great source of text- such as Wikipedia, Baglapedia, Newspapers, Magazines and etc.
- **STT**: Recorded speech can also be transformed into electronic texts using speech to text tools. This kind of component will help much in building a collection of texts of oral form.

In our work, primarily we have collected the data from different web-domains (online newspaper, Wikipedia, Banglapedia etc.) using self-made web-crawler tools. The collected data mainly represent the written aspects of the Bangla language. However, according to the original plan, we're about to include spoken corpus and scale up the current corpus to the targeted size. For collecting data from different websites we developed and used a web-crawler that can detect the targeted content and fetch it.

B. Encoding Adjustment

It's needed to be assured that, all the collected texts are in UTF-8 (Unicode) format prior to proceeding further in building this corpus. If any of the text segments is found written in non UTF-8 then, these must be converted back into Unicode. During the text collection phase, we have found that not all the Bangla text data available online are in UTF-8 format. There are still some ANSI encoded Bangla texts available on the web for legacy reasons. To solve this problem, we have developed an encode-adjusting tool that looks for encoding issues across the collected texts and adjusts and convert encodings while required.

C. Filtering

The collected text must be filtered for any unwanted, unrecognized, foreign language, misspelt words and garbage characters. Filtering can be done automatically by developing tools specifically designed for Bangla language.

Primarily we have taken care of the unwanted characters, symbols and spacing issues persisting in the electronic texts using a home-developed tool. However, due to lack of an advanced spell checker, we couldn't check the spellings of the texts. In fact, in our current scope, we do not intend to check spellings as it is just a written corpus for now.

D. Word Segmentation & Tokenizing

The next big step after filtering is segmentation/tokenizing. The process of segmenting running text into words and sentences is called tokenizing. For languages like Bangla where word segmentation can be performed by a simple script given white-space and punctuation, but still, it doesn't guarantee a 100 percent success. A tokenizer capable of handling as many as linguistically ambiguous features can only be accepted here. A token has to be linguistically significant and Methodologically useful.

In our work, we have developed a beginner level tokenizer that can break a running sentence into word forms which were later labelled by the annotator.

E. Annotation (Tagging)

We'll be using the universal format of CoNLL-U for annotation purpose. In CoNLL-U format, annotations are encoded in plain text files (UTF-8, using only the LF character as line break) with three types of lines:

1. Word lines containing the annotation of a word/token in 10 fields separated by single tab characters. The fields are namely- ID, FORM, LEMMA, UPOSTAG, XPOSTAG, FEATS, HEAD, DEPREL, DEPS, MISC
2. Blank lines marking sentence boundaries.
3. Comment lines starting with a hash (#).

Example of annotating a Bangla sentence using CoNLL-U format:

newdoc id = Rabindra_cd_20170926063000_BN

sent_id = Rabindra_cd_20170926063000_BN-0001

text = রাজেশ স্কুলে যায়।

1 রাজেশ রাজেশ PROPN NNP Number=Sing 0 root _ _

2 স্কুলে স্কুল NOUN NN Number=Sing 1 obl _ _

3 যায় যায় VERB VBZ Mood=Ind|Tense=Present 1 _ _

4 । । PUNCT । _ 1 punct _ _

V. TOOLS TO UTILIZE THIS CORPUS

We have developed some corpus analyzer tools of our own as there are very few resources available in this segment. Very few of the tools available nowadays support Bangla language. We have developed a frequency analyzer, N-grams (lexical bundles), concordance (node, KWIC, sorting, expanded context).

VI. RESULT AND ANALYSIS

Following are some of the results that were analyzed by the tools that we have developed. We have separated our corpus in 4 different plain text files of different sizes without compromising any of the qualitative features of the corpus like text-domain and other text qualities. Four parts of the plain text containing files were created in this separation process namely- mini, kilo, mega, giga. The reason behind such segmentation of the corpus file was that we wanted to make sure the corpus is easily manageable and scalable.

A. Data structure

Our primary analysis suggests that the 4 documents contain a number of 7,678,597 total words (tokens) while all the documents combined hold a total of and 285,496 unique word forms (types). The weighted average of Type-Token Ratio all the corpus is 0.0372

TABLE III. WORD TYPES AND DISTRIBUTIONS IN THE CORPUS (4 FILES)

File	Words	Types	Ratio	Word/sentence
Mini	445868	45254	0.10149	14.269145838
Kilo	756241	62095	0.08211	14.262239740
Mega	2328455	137241	0.05894	13.801686938
Giga	4148033	180135	0.04342	14.211335402

Document Length:

Longest: giga (4148033 words); mega (2328455 words)
Shortest: mini (445868 words); kilo (756241 words)

B. Word frequency

It's known that, the most frequent words in a written corpus are usually the stop words. Stop words are generally filtered out in many applications of NLP and other studies. However, here we have considered all the varieties of lexical items while preparing the word frequency list. The following table shows a frequency analysis of the lexical items that persist in the corpus.

TABLE IV. MOST FREQUENT WORDS IN THE CORPUS

Word	frequency	%	word	frequency	%
ও	74865	0.97498280	এই	27406	0.35691416
এ	51757	0.67404241	বলেন	23384	0.30453480
না	51418	0.66962754	তিনি	22981	0.29928645
করে	50955	0.66359779	এবং	22596	0.29427251
থেকে	39744	0.51759456	নিয়ে	22416	0.29192833
হয়	34932	0.45492686	এর	21860	0.28468742
করা	34615	0.45079850	হচ্ছে	21447	0.27930884
হবে	29297	0.38154105	এক	21220	0.27635257
হয়েছে	28015	0.36484530	করতে	21186	0.27590978
জন্য	27663	0.36026113	মন্তব্য	20931	0.27258886

C. Type-Token Ratio (TTR)

The ratio of the total number of words (token) in a document to the number of unique words (types) in the document is called Type-Token Ratio.

Highest:
mini (0.101)
kilo (0.082)
Lowest:
giga (0.043)
mega (0.059)

A lower vocabulary usually density indicates complex text with a pool of unique words, and a higher ratio indicates simpler text with words reused. The data indicates that, the file mini and kilo contain more 'function words' in regard to unique or content words than their siblings'-giga and mega.

Average Words per Sentence: In our corpus, we have found that the weighted average of words per sentence in our corpus is: 14.1. Below is the file specific average word per sentence rate

Highest:
mini (14.3)
kilo (14.3)
Lowest:
giga (14.2)
mega (13.8)

D. Collocation (N-gram analysis)

We have analyzed most co-occurring words or words cluster known as collocation using N-gram architecture. Below are some of the discovered collocation data of the corpus which was measured using different N-gram techniques (uni-gram and trigram).

TABLE V. THE COLLOCATION OF THE WORDS IN THE CORPUS (UNI-GRAM)

Word	count	collocate	count	word	count	collocate	count
করা	34615	হয়	7469	করা	34615	হচ্ছে	1808
করা	34615	হয়েছে	7204	হয়	34932	না	1670
এ	51757	ছাড়া	3457	এ	51757	ধরনের	1649
এ	51757	সময়	3232	না	51418	থাকলে	1541
করা	34615	হবে	3034	হয়	34932	এ	1526
হবে	29297	না	2499	এ	51757	বিষয়ে	1500
এ	51757	ব্যাপারে	2197	এ	51757	জন্য	1354
করা	34615	হচ্ছে	1808	এ	51757	কথা	1284
হয়	34932	না	1670	হবে	29297	এর	1249
এ	51757	ধরনের	1649	হয়েছে	28015	এ	1239

TABLE VI. COLLOCATION OF THE WORDS IN THE CORPUS (TRI-GRAM)

Word	count	collo-cate	count	word	count	collo-cate	count
করা	34615	হয়	7736	এ	51757	করা	2493
করা	34615	হয়েছে	7431	এ	51757	ব্যাপারে	2297
এ	51757	ছাড়া	3517	হয়	34932	না	2210
করা	34615	হবে	3505	হয়েছে	28015	এ	2201
এ	51757	সময়	3496	এ	51757	জন্য	2095
হবে	29297	না	3103	না	51418	করতে	2088
হয়	34932	এ	2955	না	51418	করা	2073
না	51418	কোনো	2737	করা	34615	না	2030
করে	50955	এ	2670	না	51418	না	1996
করা	34615	এ	2581	করে	50955	থেকে	1977

E. Data visualization

We have analyzed the data using many other techniques and tools available and developed by us and now we are to visualize some of the aspects of the corpus. Here are some examples of comparative corpus data visualization across multiple corpus data files.

Relative frequency: To find the relative frequency of any lexical item in our corpus, we need to divide the frequency of the lexical item by the total number of lexical items in the sample. In our case, the samples are the 4 separated data files of the corpus. The following chart shows the relative frequencies of the most frequent words across 4 different corpus data files.

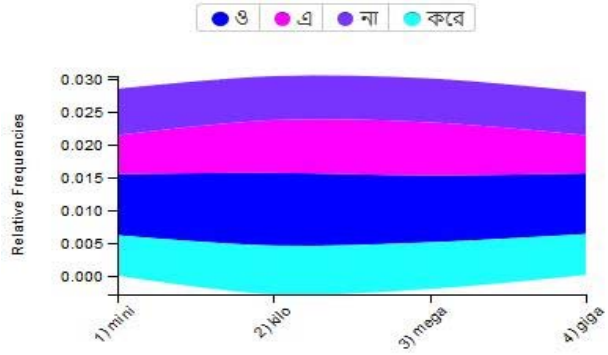

Fig. 3. Relative frequency of the top 4 (most frequent) words

F. Grammatical analysis:

We wanted to use the corpus for some more linguistic (traditional grammatical) researches as shown in Fig. 3. Therefore, we observed the comparative frequency of some of '*অব্যয়*' (which is a part of speech or grammatical category name in Bangla grammar). In comparison to English grammar, '*অব্যয়*' can be used as both prepositions, conjunction and interjection in a sentence of Bangla language. 'ও' and 'এবং' are a somewhat similar type of POS in Bangla language considering their semantic boundary and are used as a conjunction. We wanted to see how frequent are these two words and which one is more frequent than the other in Bangla language (in the context of our corpus). Below is the graph showing the result in Fig 4.

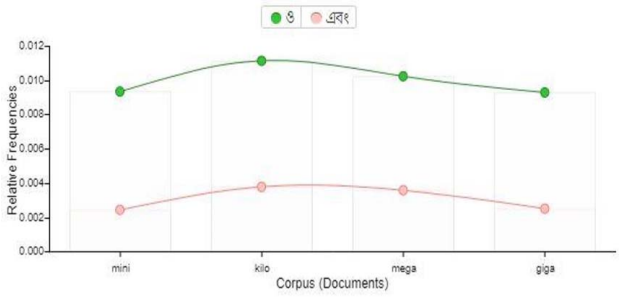

Fig. 4. relative frequency of Bangla 'অব্যয়'- ('ও', 'এবং')

CONCLUSION

The development of a corpus in our targeted scale is not only a huge task but also a tiring and a resource-hungry job. However, still, we have compiled a corpus having a size of over 7.6 million in size. Due to limitation of time and resources, we could not annotate the entire corpus with the full features that we have primarily expected. In the future, we are going to annotate the entire corpus and scale up the size of the existing corpus. Therefore we will start developing the parallel corpus shortly.

REFERENCES

[1] Gerrit Botha and Etienne Barnard, 2005. Two approaches to gathering text corpora from the World Wide Web, Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa.

[2] Salam, K. M. A., Yamada, S., & Nishino, T. (2012, May). Developing the first balanced corpus for Bangla language. In Informatics, Electronics & Vision (ICIEV), 2012 International Conference on (pp. 1081). IEEE.

[3] Salam, K. M. A., Yamada, S. and Nishino, T. 2010. "English-Bengali Parallel Corpus: A Proposal", Tokyo, TriSAI – 2010

[4] Salam, K. M. A., Yamada, S., Nishino, T. Mumit Khan, 2009 "Example-Based English-Bengali Machine Translation Using WordNet", Tokyo, TriSAI – 2009

[5] Tony McEnery and Andrew Wilson, 1996. Corpus Linguistics, Edinburgh University Press.

[6] Yeasir Arafat, Md. Zahurul Islam and Mumit Khan, 2006. Analysis and Observations From a Bangla news corpus, Proc. of 9th International Conference on Computer and Information Technology, Dhaka, Bangladesh.

[7] Baker, Mona (1995) "Corpora in translation studies: an overview and some suggestions for future research" Target 7, 2, pp 223-243.

[8] Biber, Douglas (1993) "Representativeness in corpus design", in Literary and Linguistic Computing, 8, pp 243-257.

[9] Chen, Kehjiann, Chu-ren Huang, Li-ping Chang and Hui-li Hsu. 1996. SINICA CORPUS: Design methodology for balanced corpora. Language, Information and Computation 11:167-176.

[10] Dash, Niladri Sekhar and Chaudhuri, B.B. 2001. A corpus-based study of the Bengali language. Indian Journal of Linguistics. Vol.20. No.1. Pp. 19-40.

[11] Dewan Shahriar Hossain Pavel, Asif Iqbal Sarkar and Mumit Khan, 2006. A Proposed Automated Extraction Procedure of Bangla Text for Corpus Creation in Unicode, Proc. International Conference on Computer Processing of Bengali.

[12] Frankenberg-Garcia, A. and Santos, D. (2003) "Introducing COMPARA: the Portuguese-English Parallel Corpus", Corpora in translator education, Citeseer pp 71—87.

[13] Zanettin, F. (2011). Translation and corpus design.

[14] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Sumono: A representative modern bengali corpus," SUST Journal of Science and Technology, vol. 21, pp. 78–86, 2014.

[15] S. Khan, A. Ferdousi, and M. A. Sobhan, "Creation and analysis of a new bangla text corpus bdnc01," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 5, 2017.

[16] N. S. Dash, B. B. Chaudhuri, P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja, "Corpus-based empirical analysis of form, function and frequency of characters used in bangla," in Published in Rayson, P., Wilson, A., McEnery, T., Hardie, A., and Khoja, S.,(eds.) Special issue of the Proceedings of the Corpus Linguistics 2001 Conference, Lancaster: Lancaster University Press. UK, vol. 13, 2001, pp. 144.

[17] A. I. Sarkar, D. S. H. Pavel, and M. Khan, "Automatic bangla corpus creation," BRAC University, Tech. Rep., 2007.

[18] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Supara: A balanced english-bengali parallel corpus," 2012.

[19] K. M. Majumder and Y. Arafat, "Analysis of and observations from a bangla news corpus," 2006.

[20] J. Shamshed and S. M. Karim, "A novel bangla text corpus building method for efficient information

retrieval," Journal of Convergence Information Technology, vol. 1, no. 1, pp. 36–40, 2010.

[21] Arora, S., Arora, K. K., Roy, M. K., Agrawal, S. S., & Murthy, B. K. (2016). Collaborative Speech Data Acquisition for Under Resourced Languages through Crowdsourcing. Procedia Computer Science, 81, 37-44.

[22] Nowshin, N., Ritu, Z. S., & Ismail, S. (2018, December). A Crowd-Source Based Corpus on Bangla to English Translation. In 2018 21st International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

[23] Salm, K. M., Salam, A., Khan, M., & Nishino, T. (2009). Example based English-Bengali machine translation using WordNet.

[24] Khan, M. A. S., Uchida, H., & Nishino, T. (2011, November). How to develop universal vocabularies using automatic generation of the meaning of each word. 7th International Conference on Natural Language Processing and Knowledge Engineering. IEEE.

[25] Salam, K. M. A., Yamada, S., & Nishino, T. (2011). Example-based machine translation for low-resource language using chunk-string templates. 13th Machine Translation Summit, Xiamen, China.

[26] Salam, K. M. A., Yamada, S., & Nishino, T. (2013). How to translate unknown words for English to Bangla Machine Translation using transliteration. Journal of computers, 8(5), 1167-1174.

[27] Salam, K. M. A., Uchida, H., Yamada, S., & Nishino, T. (2012, August). UNL Ontology Visualization for Web. In 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (pp. 542-545). IEEE.

[28] Salam, K. M. A., Uchida, H., & Nishino, T. (2012, December). Multilingual universal word explanation generation from unl ontology. In 24th International Conference on Computational Linguistics (p. 137).

[29] Salam, K. M. A., Setsuo, Y., & Nishino, T. (2011, December). Translating unknown words using WordNet and IPA-based-transliteration. In 14th International Conference on Computer and Information Technology (ICCIT 2011) (pp. 481-486). IEEE.

[30] Uchida, H., Zhu, M., & Khan, M. A. S. (2012, December). UNL explorer. In Proceedings of COLING 2012: Demonstration Papers (pp. 453-458).

[31] Salam, K. M. A., Uchida, H., Yamada, S., & Nishino, T. (2013). Web Based UNL Ontology Visualization. Journal of Convergence Information Technology, 8(13), 69.

[32] Salam, K. M. A., Setsuo, Y., & Tetsuro, N. (2012, December). Sublexical Translations for Low-Resource Language. In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (pp. 39).

[33] Salam, K., Yamada, S., & Tetsuro, N. (2012). Phonetic Bengali Input Method for Computer and Mobile Devices. In the Proceeding of the Second Workshop on Advances in Text Input Methods (WTIM 2), COLING (pp. 73-78).

[34] Chaudhury, S., Dasgupta, S., Munawar, A., Khan, M. A. S., & Tachibana, R. (2017, September). Text to image generative model using constrained embedding space mapping. In 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1-6). IEEE.

[35] Salam, K. M. A., Setsuo, Y., & Nishino, T. Using WordNet to Handle the OOV Problem in English to Bangla Machine Translation. In GWC 2012 6th International Global Wordnet Conference (p. 35).

[36] Salam, K. M. A., Yamada, S., & Tetsuro, N. (2017, July). Improve Example-Based Machine Translation Quality for Low-Resource Language Using Ontology. In International Conference on Applied Computing and Information Technology (pp. 67-90). Springer, Cham.

[37] SALAM, K. M. A. (2014). Ontology Based Machine Translation for Bengali as Low-resource Language (Doctoral dissertation, UNIVERSITY OF ELECTRO-COMMUNICATIONS).

[38] Salam, K. M. A., Uchida, H., Yamada, S., & Nishino, T. (2013, June). Universal Words relationship question-answering from UNL Ontology. In 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS) (pp. 423-427). IEEE.

[39] Salam, K. M. A., Tetsuro, N., & Yamada, S. (2012, December). Bangla Phonetic Input Method with Foreign Words Handling. In Proceedings of the Second Workshop on Advances in Text Input Methods (pp. 73)