SUST Journal of Science and Technology, Vol. 16, No.2, 2012; P:46-51

SUPara: A Balanced English-Bengali Parallel Corpus

(Submitted: October 16, 2011**; Accepted for Publication:** May 12, 2012**)**

Md. Abdullah Al Mumin, Abu Awal Md. Shoeb, Md. Reza Selim and M. Zafar Iqbal
Dept. of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh
Email: mumin-cse@sust.edu, shoeb-cse@sust.edu, selim@sust.edu, mzi@sust.edu

Abstract

Parallel corpora have become an essential resource in natural language processing. In spite of their importance in many multi-lingual applications, a few effective English-Bengali corpus has been made available, given the scarcity of its resources and the intensive lobors required in its creation. This paper introduces Shahjalal University Parallel (SUPara) corpus, an English-Bengali sentence-aligned parallel corpus consisting of more than 200,000 words in either languages, which is the largest among freely released corpus of its kind.

A balanced corpus refers to carefully selected and fully described body of natural language texts, which more or less represent the language. SUPara is balanced according to five text types (literature, journalistic texts, instructive texts, administrative texts and texts treating external communication) and freely accessible to the research community. In this paper, we address the development process of SUPara corpus in context of its balanced design, universal encoding, lingustic markup and sentence alignment. The statistics of the corpus are also presented here. To the best of our knowledge, SUPara has been the first freely released balanced English-Bengali corpus.

Key Words: parallel corpora; corpus design; balanced corpus.

1. Introduction

A parallel corpus is a collection of text, paired with translations into another language[1]; whereas a monolingual corpus contain texts in a single language.

Parallel corpora have proved themselves to be extremely useful resources for cross-linguistic research and translation studies in recent years. They provide an empirical basis for contrastive and typological research and they give new insights into the languages compared - insights that are likely to be unnoticed in studies of monolingual corpora. They are also important for practical applications in various fields, such as lexicography, language teaching and computer-aided translation.

The value of linguistic corpora is not yet acknowledged in Bangladesh, although in recent times some sporadic attempts are made for designing monolingual corpora in Bengali. While there are various resources such as newswires, books and websites that can be used to construct monolingual corpora, parallel corpora need more specific types of multilingual resources which are comparatively more difficult to obtain. As a result, parallel corproa are rarely available especially for lesser studied languages like Bengali.

In order that a parallel corpus might be useful for researchers working in different areas of natural language processing, it has first of all to be well-balanced and should have easy access to the researchers[2]. A balanced corpus refers to carefully selected and fully described body of natural language texts, which more or less represent the language. As language is an open set, it is very challenging task to build a balanced corpus.

This paper presents SUPara corpus, an English-Bengali parallel corpus. The SUPara corpus does exhibit text type balance and is available for the entire research community, which is first of its kind for the English-Bengali pair.

The remainder of this paper is structured as follows: section 2 reviews on previous parallel corpora for Bengali. Section 3 describes the development of the SUPara corpus. Section 4 gives an overview of the statistics of the corpus. Section 5 focuses on further development of the SUPara corpus. Section 6 concludes the paper.

2. Previous Bengali Parallel Corpora

Until the release of SUPara corpus, there were quite a few parallel corpora for Bengali language. After examining these corpora, we notice that some are freely available but lacking in text type balance and that others include several text types but are not freely accessible to the research community.

EMILLE[3] corpus developed through EMILLE project (Enabling Minority Language Engineering), which was undertaken by the universities of Lancaster and Sheffield. This corpus consists of a series of monolingual corpora for fourteen South Asian Languages and a parallel corpus of English and five of these languages.

A subset of EMILLE corpus is Bengali parallel corpus, which consists of 200,000 words of text including tag words. This parallel corpus uses UK government advice leaflets as data sources. Since these leaflets cover a wide range of areas, EMILLE corpus exhibits balancedness at some extent. However, data taken only from one domain, that is, UK government advice leaflets, undermine its balancedness. The major drawback of EMILLE corpus is that the corpus is not sentence alligned. The corpus has been made alligned at document level only.

EMILLE corpus is publicly available for research purposes. However, one should go through some paperworks to get the corpus officially. Table 1 depicts a comparative picture between SUPara corpus and EMILLE corpus when no tag words are included. Lexical diversity score shows that on average each vocabulary item appears less frequently in SUPara corpus than in EMILLE corpus for both English and Bengali sides.

	SUPara		EMILLE	
	English	Bengali	English	Bengali
Corpus size (in words)	244539	202866	170758	172091
Vocabulary size (no. of unique words)	14571	22456	9222	16970
Lexical diversity	16.78	9.03	18.52	10.14

Table 1: Comparison between SUPara and EMILLE corpus

OPUS[4] is a freely available corpus, which consists of parallel texts in 60 languages. This corpus uses open source documentation and their translation as source. For Bengali parallel corpus, OPUS only uses KDE4 manual texts. Due to lack of varieties of text type OPUS corpus exhibits severe text type imbalance.

The authors in [5] describe the development process of the first balanced English-Bengali parallel corpus. They propose a corpus selection criteria in the context of low resource language like Bengali to develop a balanced corpus. The target size of the parallel corpus is more than 10 million words. The resulting corpus, however, has not yet been released.

3. Development of the Corpus

It is now a well recognized fact that a corpus is more than just a collection of electronic texts. Corpus data have to be selected with care with respect to the intended applications. In this project we emphasize quality with regard to content and translation. We focus on a collection of written text to build a balanced corpus of the source and target language.

3.1 Data Sources

We have used texts from the following sources that are either publicly available or granted permission from respective copyright holders.

- A novel 'আমার বন্ধু রাশেদ' (Rashed, My Friend) and a feature 'মুক্তিযুদ্ধের ইতিহাস' (History of the Liberation war), both are written by Muhammed Zafar Iqbal and translated by Yashim Iqbal, and has been provided by the author.

- Documents containing Prime Minister's speeches, budget speech, commercial policy, education policy, rules and procedure of parliament, press releases are obtained from government official website.
- Several essayistic texts, news reporting articles 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.
- Several documents are generated by mining data from Wikipedia and Banglapedia.

3.2 Balanced Design

The SUPara corpus considers following two features in its design:

Text type diversity

The greater text diversity in the corpus, the more universal is potential use. The corpus is balanced in the way that it offers a great variety of text materials coming from different domains divided into five major text types: literature, journalistic texts, instructive texts, administrative texts and texts treating external communication.

Corpus structure

The typology and structure of the initial design were based on the prototype approach by David Lee[6]. In order to prevent having overbroad categories containing heteregeneous material, Lee advocates using a prototype approach based on the basic-level category and thus creating a multi-level typology. This means introducing subcategories and adding this information to the metadata which allows the user to fine-tune his/her search. This approach led us to opt for a two-level typology as presented in Table 2.

3.3 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. The following steps have been applied as preprocessing on the English and Bengali documents:

SUPERORDINATE	BASIC LEVEL
1.Literature	1.Novels
	2.Essayistic texts
	3.(Auto)biographies
	4.Expository non-fictional literature
2.Journalistic texts	1.News reporting articles
	2.Comment articles (background articles, columns, editorials)
3.Instructive texts	1.Manuals
	2.Internal Legal documents
	3.Procedure descriptions
4.Administrative texts	1.Legislation
	2.Proceedings of debates
	3.Minutes of meetings
	4.Yearly reports
	5.Correspondence
	6.Official speeches
5.External Communication	1. (Self-)presentations of organizations, projects, events
	2.Informative documents of a general nature
	3.Promotion and advertising material
	4.Press releases and newsletters
	5.Scientific texts

Table 2: SUPara's two-level typology.

Cleaning up documents

We start by cleaning up the original material that we collected from the different sources. 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

We use simple principles for the encoding of documents in our corpus. The texts are encoded according to international standards by using UTF8 (Unicode). For Bengali documents we have used Nikosh[1] converter to encode all formats into Unicode.

All unicode formatted data are then marked up according to the corpus encoding standard for XML (XCES)[2] [7].

Alignment

The alignment of translated segments with source segments is essential for building parallel corpora. First of all, we have maintained alignment at document levels. Each file in a sub-directory is aligned separately with its translation to keep alignment errors at a low level.

At the sentence level, we use manual alignment at the moment. Though automatic sentence alignment is robust, however, automatic alignment approaches do not produce results with reasonable accuracy for unrelated language pairs. English and Bengali are very much unrelated language in context of sentence length, sentence order and morphological richness. Therefore, automatic alignment approaches that work better for related language may not be fruitful. We were aware about this inefficiency of automatic alignment for English-Bengali language pair from the beginning of the corpus creation. This leads us to align sentences manually rather than using automatic alignment whenever any parallel document become available. As a result the SUPara corpus provides the sentence alignment at high accuracy level, which is the strength of this corpus.

Tools

We have used various tools for preparing the corpora included in SUPara. In particular, we apply various types of open-source software and free research tools. These tools include sentence splitter, word histogram generator, unicode converter, etc. For English part of the corpus we use uplug[3] tools and NLTK library[4]. For Bengali part of the corpus we use our own developed tools[5]. They are freely available and can also be used by others to produce similar corpora.

3.4 Availability

The corpus is available free of charge for educational and research purposes, however, the license allows collecting statistical data and making short citations. The corpus will be distributed through the Computer Science and Engineering (CSE) department of Shahjalal University of Science and Technology (SUST)[6].

4. Statistics of SUPara

Table 3 summarizes the statistics[7] of the first release of SUPara. From the data in Table 3, we can draw following contrastive pictures between the two languages:

	English side	Bengali side
Corpus size (in words) excluding punctuations	244,539	202,866
Corpus size (in characters excluding spaces)	1,229,948	1,119,985
Average sentence length (in words)	11.56	9.59
Vocabulary size (no. of unique words)	14,571	22,456
Corpus size (in lines)	21,158	

Table 3: Statistics of SUPara

[1] http://www.ecs.gov.bd/nikosh

[2] http://cs.vassar.edu/XCES/

[3] http://www.sourceforge.net/projects/uplug

[4] http://www.nltk.org

[5] http://www.sust.edu/cse/corpus/tools

[6] http://www.sust.edu/cse/corpus/SUPara

[7] according to the data on October 3, 2011.

- The corpus size in English side is usually larger than that of the Bengali side. This is due to the English function words that occur more frequently than content words in English text. Content words refer to objects, actions, or properties and function words tell us how these content words relate to each other. When translated, most of these English function words are used together with Bengali content words, thus word count falls down in Bengali text. For example, consider the English text 'The boy go to the school' and its translation into Bengali text 'ছেলেটি স্কুলটিতে যায়'. The English function words 'the' and 'to the' are translated into 'টি' and 'টিতে', respectively and use with Bengali word 'স্কুল'. Thus, Bengali word count falls into 3 compare to the English word count 6. Table 4 displays the 10 most frequent words used in the SUPara corpus for both sides.

English side			Bengali side		
Words	Frequency	(%)	Words	Frequency	(%)
the	14596	5.97	ও	2922	1.44
of	8959	3.66	এবং	2364	1.17
to	7270	2.97	করা	2032	1.00
and	6895	2.82	না	1976	0.97
in	4430	1.81	জন্য	1900	0.94
a	3678	1.50	হবে	1800	0.89
be	3418	1.40	করে	1721	0.85
for	2825	1.16	এই	1320	0.65
will	2210	0.90	কোন	1256	0.62
is	1947	0.80	বা	1151	0.57

Table 4: The top 10 frequent words in the SUPara corpus

- On the other hand, the vocabulary size in Bengali side of the corpus is about 1.6 times more than that of English. This is due to the rich inflectional morphology of Bengali language. For example, consider two english texts - 'I eat' and 'You eat' and its translation into Bengali – 'আমি খাই' and 'তুমি খাও', respectively. We get two morphological forms 'খাই' and 'খাও' of english word 'eat' when translated into Bengali due to the difference of 'Person' number at the subject. Thus, vocabulary count increases to 4 in Bengali side compare to the vocabulary count 3 in English side for these two texts.
- The Bengali sentences are on average constructed using fewer characters in comparison to their equivalent English sentences.

SUPara corpus contains altogether 80 document pairs. Full details on the corpus size are given in Table 5.

	Document pairs	Sentences	Words		Lexical diversity	
			English	Bengali	English	Bengali
Literature	2	4,974 23.51%	43,916 17.96%	37,503 18.49%	11.61	6.79
Journalistic Texts	14	2,746 12.98%	30,878 12.63%	24,861 12.25%	6.22	3.76
Instructive Texts	2	3,074 14.53%	39,508 16.15%	31,470 15.52%	13.40	9.00
Administrative Texts	16	6,333 29.93%	79,684 32.59%	65,149 32.11%	10.83	6.34
External Communication	46	4,031 19.05%	50,553 20.67%	43,883 21.63%	9.04	6.40
Total	80	21,158 100%	244,539 100%	202,866 100%	16.78	9.03

Table 5: SUPara sections and data sizes

5. Further Development

In future, we plan to further enlarge SUPara corpus. Even now we have a lot of various collections of parallel material available in queue to process and merge into SUPara. We also intend to develop appropriate tools that support preprocessing of Bengali parallel corpora, for example, automatic sentence aligner for English-Bengali parallel corpora.

We would like to annotate SUPara corpus on various levels up to deep syntactic layer. We also plan to designate subsections of SUPara as standard development and evaluation data sets for machine translation, paying proper attention to cleaning up of these sets. Our future plans include experimenting with several machine translation systems.

6. Conclusion

In this paper, we have presented the SUPara corpus, a collection of Engish-Bengali parallel texts. In its design, SUPara is made balance in respect of variety of text material from different domain and is opted for a two-level typology structure. We have passed through several preprocessing steps to make SUPara functional. The texts in various formats are normalized into plain texts. Then these plain texts are encoded in Unicode format and marked up according to the corpus encoding standard for XML (XCES). For retaining high accuracy, we have aligned the corpus at document level and sentence level manually.

We have presented statistics of SUPara which help us to analyze some contrastive features between English and Bengali languages. The corpus which consists of 244,539 words in English and 202,866 words in Bengali is the largest available English-Bengali parallel corpus. It is available free of charge for educational and research purposes. To our knowledge, it is the first balanced and publicly available parallel corpus for the English-Bengali pair.

The corpus is still under development and we hope to enlarge it further to achieve a large-scale (in order of million words) parallel corpus. We also hope that the SUPara corpus will initiate all the research in multilingual NLP applications for Bengali that did not become possible due to lack of a publicly available balanced parallel corpus.

References

[1] Koehn, P., 2010. *Statistical Machine Translation Book.* page 54. Cambridge University Press.

[2] Rura, L., Vandeweghe W. and Montero Perez, M., 2008. *Designing a parallel corpus as multifunctional translator's aid.* In Proceedings of the 18th FIT world congress, Shanghai, China.

[3] McEnery, A., Baker, P., Gaizauskas, R. and Cunningham, H., 2001. *EMILLE: building a corpus of South Asian Languages.* In D. Lewis and R. Mitkov (eds) Proceedings of Machine Translation and Multilingual Applications in the New Millennium.

[4] Tiedmann, J. and Nygaard, L., 2004. *The OPUS corpus - parallel and free.* In Proceedings of the Fourth International Conference on Language Resources and Evaluation(LREC'04), Lisbon, Portugal, pages 93-96.

[5] Salam, K.M.A., Setsuo, Y. and Nishino, T. *English-Bengali Parallel Corpus: A Proposal.* TriSAI-2010, pp242, Beijing, China, October 2010

[6] Lee, D.Y.W., 2001. *Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle.* In Language Learning and Technology, 5(3): 37-72.

[7] Ide, N. and Priest-Dorman, G., 2000. *Corpus Encoding Standard - Document CES 1.* Technical Report, Dept. of Computer Scince, Vassar College, USA and Equipe Language et Dialogue, France.

[8] Bojar, O. and Žabokrtsky, Z., 2006. *CzEng: Czech-English Parallel Corpus, Release version 0.5.* Prague Bulletin of Mathematical Linguistics, 86. pp 59-62.

[9] Gale, W.A. and Church, K.W., 1993. *A program for aligning sentences in bilingual copora.* Computational Linguistics, 19(1): 75-102.

[10] Koehn, P., 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation.* In proceedings of the Tenth Machine Translation Summit, Phuket, Thailand, pages 79-86.

[11] McEnery, A. and Wilson, A., 1996. *Corpus Lingistics.* Edinburgh: Edinburgh University Press.