Bangla Web Corpus: Crawling in the Web and Fishing with the Net

Niladri Sekhar Dash, Devika Shukla and **Sayantani Pathak**
Linguistic Research Unit, Indian Statistical Institute, Kolkata, India
Email: ns_dash@yahoo.com

Abstract

This paper makes an attempt to describe and discuss the process of development of a new Bangla monolingual digital text corpus (namely, the Bangla Web Corpus (BWC)), which is developed as a part of the ILCI-2 project supported by DietY, Govt. of India with textual data retrieved from internet, digital portals, and web pages. It also tries to address the methods and strategies that are applied for this purpose; the issues that have cropped up in the act of generating the whole corpus database; and the major problems that are faced at the time of creating the corpus. In our opinion, the issues that have cropped up in the process, the problems that are faced, and the strategies and methods that are adopted to achieve the goal can give clear insights to deal with similar situations for generating corpora in other less resourced and less computer-savvy Indian languages. The acts of fishing language data from the web and harvesting the BWC may be treated as milestones in the history of Bangla corpus generation, as the BWC holds tremendous potentials for opening up new avenues for web crawling and language corpus building in the wider spectrum of research in language technology and applied linguistics. An on-line version of the BWC that is on the verge of being hoisted in the net, will contribute towards building an interface where language users are allowed to navigate through web-enabled corpus to address their linguistic needs. Here lies the theoretical relevance, empirical pertinence, and functional importance of the work which seeks to propose a makeshift guideline for the new generation of corpus developers in Indian languages.

Keywords: monolingual corpus, language technology, Unicode, Bangla, data collection, text types, text domains, digital texts, metadata

1. Introduction

A language corpus, at present, is extensively used in all major areas of descriptive linguistics, applied linguistics, and language technology, as a language corpus, due to its composition with actual collection of empirical language use has been accepted as an authentic source of linguistic data, information, and examples. The term 'corpus', in principle, signifies that it is capable of representing potentially unlimited selections of text of a language (Dash 2015: 4). That means a balanced and multi-disciplinary language corpus is adequately representative of a given language or a variety to make it maximally useful for all kinds of linguistic study and application.

A monolingual corpus— an important type of a corpus— contains a large collection of text samples derived from a single language representing the use of the language in various fields

of linguistic activity. The homogeneity of structure, diversity of text types, and uniformity in text representation are the rudiments of a monolingual corpus, which becomes indispensible in compilation of generalised lexicon as well as for language description and analysis. Usually a monolingual corpus is designed following some predefined criteria that are normally used for generating a reference corpus (Dash 2005: 15) because a monolingual corpus, in its core, usually aims at representing a language in a general fashion.

Keeping this attribute in mind, attempt is made to develop a Bangla Web Corpus (BWC) as a part of the Indian Languages Corpora Initiative-II (ILCI-II) under the aegis of Technology Development for the Indian Languages (TDIL) initiated by the DietY of the Ministry of Communication and Information Technology (MCIT), Government of India. The purposes behind the development of this corpus may be visualized in direct utilization of this resource in development of language processing tools like part-of-speech tagger, spelling checker, chunker, lexical collocator, morphological analyser, parser, lemmatizer, text editor, named entity identifier, etc. The BWC may also be visualised to be a useful resource for compiling monolingual generalized lexical databases, termbanks, and function wordlists.

Within a specific research frame, the task of generating domain-specific monolingual corpus has many limitations because the freedom for collection of data across all spatio-temporal boundaries are often sealed due to limited scope of the project. In principle, the predefined guidelines and target of the project often dictates a corpus developer to fold his/her wings within the nest built for the purpose. The present scheme of the project does not go beyond this norm as our task is ear-marked to a specific goal of generating 50000 sentences across various domains and sub-domains of language use in digital net. Moreover, the corpus collected in this manner also meant to be validated and processed to make it maximally user-friendly in the world of language computation. Although the path is defined and mission is visualized, the actual journey was plagued with many meanders and mirages, that are highlighted in the subsequent sections. However, before this, it is necessary to have some theoretical knowledge about the form, nature, and content of a web corpus, which may help to understand how a web corpus differs from that of a general corpus both in composition and application.

This paper tries to showcase experiences relating to development of a new monolingual Bangla Web Corpus (BWC). In Section 2, it concentrates on the features and content of a web corpus; in Section 3, it discusses the purposes behind this corpus generation; in Section 4, it refers to some early attempts made to compile web corpus in other (mostly Non-Indian) languages; in Section 5, it refers to the methodologies applied to create the web corpus; in Section 6, it describes in brief the metadata information tagged to the corpus; in Section 7, it focuses on the problems faced during the course of the work; and in Section 8, it identifies the utility of a web corpus in various domains of linguistics and language technology.

2. Defining a Web Corpus

A World Wide Web (WWW), as a source of electronic language data, is gaining popularity quite rapidly because it had opened Pandora's Box for language corpus having features like enormous size, huge content, wide variety, broad linguistic dimension, geographical diversity, wide register variation, demographic difference, up-to-date status, varied text types,

synchronic expanse, diachronic range, and multimodal texture, etc. A web corpus with all these (and more) features is obviously a much better linguistic resource than a corpus made with printed text samples for the people working in various domains of linguistics and language technology.

A web corpus, in principle, is different from a standard digital corpus on many aspects relating to its content and composition. While a normal digital corpus contains text samples from both physical and virtual worlds (the balance is normally tilted towards text samples obtained from physical world), a web corpus, solely and wholly, unless otherwise desired, contains texts of the virtual world. That means the language data that are used to build up a web corpus, are practically and primarily collected from various web sites, homepages, and similar virtual sources only. The texts samples are totally digital in nature as no printed text is typed in to be included in it.

Moreover, dissimilar to multimodal corpus, a web corpus does contain neither imaged text data (i.e., texts in pdf, JPEG format, etc), nor encoded data (i.e., HTML, SGML, TEI, etc.). Also, it is free from all kinds of non-textual elements and properties (i.e., diagrams, tables, charts, pictures, animations, graphs, flowcharts etc.) as such elements may hamper, in later stages, the activities of corpus processing and linguistic data retrieval from the corpus. Besides, the text samples are completely Unicode compatible therefore, globally accessible in all formats and platforms of NLP works. Furthermore, the nature of text is both formal and informal since the samples are compiled from official sites, personal blogs, and social networking sites. Due to these factors the language of the web corpus is both personal and public, informative and imaginative, casual and careful, and well-formed and ill-formed. This gives a web corpus a unique linguistic identity, which is different from other digital text corpora as well as from printed text corpora. It is a class of its own, which has tremendous potentials to reflect on the varied texture and colourful fabric of the language in use in the cyber world. Based on its form, formation, content and composition it is very much possible to identify some notable characteristic features of a web corpus in the following manner:

[1] It contains large amount of data of actual language use in the virtual world.
[2] It contains diverse text types. Texts are collected from diverse sources.
[3] It captures varied in spatio-temporal features of language use in cyber world.
[4] The language data captured in it is both synchronic and diachronic in nature.
[5] The structure of a web corpus is usually balanced in composition. It may, however, be skewed, if a particular research works desires so.
[6] A web corpus is adequately representative of the present state of the target language from which the corpus is developed. Since its goal is to represent the present state of the language in use in question, it normally tries to be representative as far as it possible with wide variety of text types from different domains of language use.
[7] The texts included in a web corpus are mostly un-annotated in form. Similar to general corpus the texts are stored in its raw form, with provision for extra-textual and intra-textual annotation — whenever required — with a scope for returning back to the original raw text.
[8] A web corpus contains both formal and informal texts - as the source of data is open web sources. Formal texts from official sites as well as informal texts from personal and social sites contribute to the constitution of the corpus.

[9] Easy augmentability is a unique feature of a web corpus. As and when required it can be updated with new sets of data to overcome paucity of data and to overcome skewedness and imbalance in text representation.

[10] The texts samples of a web corpus, similar to that of a general corpus, is always open for verification and validation. Since texts are obtained from freely accessible web sites and homepages, anyone can verify, at any time, the validity of the text data just by referring to sources of the data.

[11] Since text samples are representative collection of the actual language in use, the corpus is maximally authentic with regard to the originality of the text as well as with regard to the present state of the language. In fact, authenticity of text samples in the web corpus is beyond doubt as it is faithfully depicts the present state of the language in question.

[12] Quick and repeated retrieval of linguistic data, information and examples, similar to a general corpus, is an important feature of a web corpus. Data can be extracted from this corpus quite easily and for this task one does not need to be an expert in computer use. Also, additional skill is not required to refer to the source sites.

[13] Due to addition of the metadata to the original text samples, it is easy to process the data stored in the web corpus. All corpus processing techniques (such as, lexical sorting, frequency count, concordance, local word grouping, morphological processing, lemmatization, collocation, lexical categorisation, compound decomposition, POS tagging, chunking, parsing, named entity identification, anaphora marking, etc.) can be easily applied on a web corpus to make it maximally usable in all kinds of linguistic works.

[14] Since the texts in a web corpus are available in Unicode format, the texts are maximally computable. Texts can be used in all kinds of computational platforms and processing interfaces irrespective to any font and orthographic uniqueness or variety.

[15] Finally, the text database is always available for customization — a major advantage of a web corpus. Based on specific requirement the database can be minimized, curtailed, shortened, compressed, deleted and customised to fit into the frame of individual research requirements.

Careful consideration of the feature mentioned above clearly indicates that a web corpus is a unique type of corpus, which can have many advantages over a traditional language database or a written text corpus. In fact, it will not be an utopian expectation if someone visualizes the ever increasing use of web corpus in the mainframe language research and application within a few years to come.

3. Purpose behind the Bangla Web Corpus

The task of BWC generation is initiated as a part of the research project, namely, *Indian Languages Corpora Initiative-2* (ILCI-2) under the banner of *Technology Development for Indian Languages* (TDIL), with full financial support of the DietY, MCIT, Govt. of India. The primary agendas of this project are as follows:

(a) Generation of domain-specific parallel translation corpora with Hindi as the source language and other Indian languages (including English) as the target languages.

(b) Generation of multi-disciplinary monolingual corpus with web-based texts in all major Indian languages (alphabetically: Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Konkani, Malayalam, Marathi, Nepali, Odia, Punjabi, Tamil, Telugu, and Urdu).

In simple terms, the primary goals stated above in a poly-lingual country like India are simply indispensible as the country with 4 (or 5) language families, 22 scheduled national languages, and more than 1000 unscheduled languages (as mentioned in Census 2011) is waiting for such innovative projects to benefit its languages and people. With regard to the second agenda of the ILCI-2 project, the major purposes behind building the monolingual web-based corpus may be summarised in the following way:

(a) The data of BWC is meant to be used for developing language processing tools for POS tagging, spelling checking, lexical collocation, word processing, morphological analysis, parsing, concordance, lemmatisation, text editing, etc.
(b) BWC can be used to develop domain specific digital lexical database for each Indian languages.
(c) BWC can be used to compile domain specific monolingual dictionaries as well as machine readable dictionaries.
(d) BWC can also be utilised to develop translation support system, language resource system as well as information retrieval system.
(e) BWC can be used to design web based learning system in Indian languages.
(f) BWC is useful for theoretical linguistic studies, such as, language and subject domains, language change across domains, patterns of semantic change of words, ambiguity in words, structure of sentences across text types, knowledge representation through text formation, information embedding in texts types, etc.
(g) BWC, due to its composition with text of different subject areas, domains, genres, and fields, is informative and useful for various linguistic and non-linguistics studies, cross linguistic comparison, and other works of descriptive and applied linguistics.
(h) BWC is also useful for compilation of domain-specific technical terms, scientific words, phrases, set expressions, neologism, idiomatic tracks, and proverbial strings, etc. which are necessary for understanding the present state of a languages as well as for preparing language teaching texts and study materials.
(i) BWC may be used as a primary source to look into the nature of sense variation of words in present day language use - thereby designing a network of sense variation of words to be adequately represented in the lexical profile of the existing WordNet.

This shows that a monolingual web corpus, like a general corpus, is extremely useful in many works of language and linguistics. Therefore, an innovative project like the above can put great resource in the hands of Govt. of India to make useful language policies and planning for better utilization of privileged Indian languages as well as for preservation and promotion of the under-privileged languages.

4. Early Attempts for Web Corpus Generation

The history of language corpora generation in electronic form is more than half a century long. Starting with the Brown Corpus (Francis and Kucera 1964, Kucera and Francis 1967), over the decades, we have travelled a long distance and in this long journey we have come

across many electronic text corpora of different types, texts, forms, contents and compositions designed with different corpus design criteria (Atkins et al. 1992). Although it is not required to roam on this history peeping into every corner of the landscape (Dash 2008), the referential relevance of these corpora in the context of web corpus compilation cannot be ignored.

This journey (Dash 2009), however, does not reflect on the event of web corpus generation as this is a very recent phenomenon, which is trying to capture our attention with a tantalizing invitation for exploration into all its trenches and treasures. In this present section, we hope to present a short sketch of web corpus panorama as this genre is yet to flourish into its full shape.

The first effort, as far as we know, is made as a strategy to produce some language corpora with texts from the internet, which may be used as a joint linguistic resource (along with electronic text corpus) for various linguistic activities (Bergh et al. 1998). This leads some corpus developers to make a quantum leap from the British National Corpus (BNC) to the 'cyber corpus' keeping in mind that internet is gradually opening up to the corpus developers with a varied universe of language data (Brekke 2000). At the same time the whole idea of treating web as a corpus rather than using web text data to build an electronic corpus has come to us as a novel concept that has the potential to give new direction in the journey of web corpus development (Kilgarriff 2001). With this new insight the problems and issues as well as the modus operandi for generating corpora from the web texts required close investigation so that least amount of error is made in the act of web corpus compilation (Cavaglia and Kilgarriff 2001).

Within years we understood that searching World Wide Web for language corpus generation (Lawrence and Giles 1998) as well as using World Wide Web itself as linguistic corpus are those tricky tracks that are not so easily traversable by one and all (Meyer *et al.* 2003). So it was necessary to design crawling system for extracting web text data for several linguistic purposes (Baroni 2005). Even then it was a real challenge to dispel the cloud of scepticism from the frozen minds with regard to generation of open-source corpora by using the internet to fish for the linguistic data, which has been a uphill task for many corpus linguists in recent years (Sharoff 2006).

Since the process of making the web more useful as a source for linguistic corpus has been an area of recent investigation and enterprise (Fletcher 2004), the issues that are involved in extracting linguistic data and information from the web to produce web corpus have been addressed quite adequately with reference to some web corpora (Renouf *et al.* 2004). The scope and utility of web corpus is further expanded when we find that attempt is already made to build web corpora for minority languages by learning to generate web search queries and internet sites (Ghani *et al.* 2003). Such expansion of scope is however, put to challenge, when we are informed that American National Corpus (ANC), which is reportedly made with electronic written and spoken texts, contains more than the web can provide (Ide *et al.* 2002).

On the other hand, we have been instructed how diachronic linguistic analysis is possible on the web corpus with application of WebCorp tool (Kehoe 2006). Also, we have been

informed that it is very much possible to use WebCorp as a two-edged tool to access web for linguistic works as well as apply linguistic data and information to access web texts (Kehoe and Renouf 2002). This leads to generation of a new corpus from the web by making web text more 'text-like' in form, content and texture (Kehoe and Gee 2007) as well as using the web text data into weaving a diachronic corpus patchwork (Kehoe and Gee 2009). The reality is that making of WebCorp — a web corpus generation tool has provided a renewable data source for the corpus linguists (Renouf 2003), as this tool has not only helped in text data compilation but also in filling the need of a search engine of a linguist to supplement existing text resources (Renouf and Kehoe (2013).

The brief sketch presented above, however unfortunate it may sounds, does not speak about the less resourced languages the spectrum in which almost all the major Indian languages fall. This is simply because we have not yet readied ourselves to explore the possibilities of using web sites made with Indian languages to collect data for generating web corpus in Indian languages. Obviously there are many technical, linguistic, legal and logistic barriers involved herein without removal of which it is really difficult to achieve success, however small, in this enterprise. The present paper, perhaps, is the first of its kind in Indian languages where we have made an attempt to present a short sketch on our effort for developing a web corpus in Bangla.

Although, a decade ago, the importance of electronic corpus of any kind in Indian languages has been explained quite elaborately (Dash 2004) with full details about the methods used for designing electronic text corpora in some Indian languages (Dash 2007), the actual effort for corpus generation in Indian languages has not been much encouraging except the recent attempt for generation of parallel translation corpora across Indian languages in the project of the DietY, Got. of India (Dash 2012). Sporadically, however, we come across information about generation of corpora in Indian languages, a recent effort of this kind is that an attempt for adopting a structured approach for building Assamese electronic corpus with data from printed and digital sources (Sharma *et al.* 2012).

5. Methodologies Applied

While generating this web corpus, we have applied various methodologies through which we could extract data in a uniform manner from various domains and sub-domains. The general issues relating to generation of a text corpus in all natural languages are also pertinent in this context. The major issues that are considered in this case with utmost importance include overall design of the web corpus, selection of domains and sub-domains of texts, range of data to be collected, process of data collection, and validation of the raw corpus, etc. which are discussed in some details in the following sub-sections. What is most striking here is that we had to face many challenges that we came across while we were generating the BWC. The challenges are of two types — linguistic and non-linguistic — and these are explained in subsequent sections. We believe that the issues that are addressed here may be considered as useful inputs for web corpus creation by others.

5.1 Overall Design of the Web Corpus

The overall design of the BWC is an important factor of serious consideration. The tool for generating this monolingual web corpus is designed in such a manner that it opens up with a useful on-line interactive interface that facilitates operations like corpus storage, text editing, and data search. Till date, a total number of 90,000 natural Bangla sentences are obtained from various genres and these are processed and uploaded in this interactive interface through online mode with the help of the corpus generation tool.

5.2 Domains and Sub-domains of Texts

The text samples that are collected following the guidelines of the ILCI-2 are distributed into eighteen (18) different domains as shown below (Fig.1):

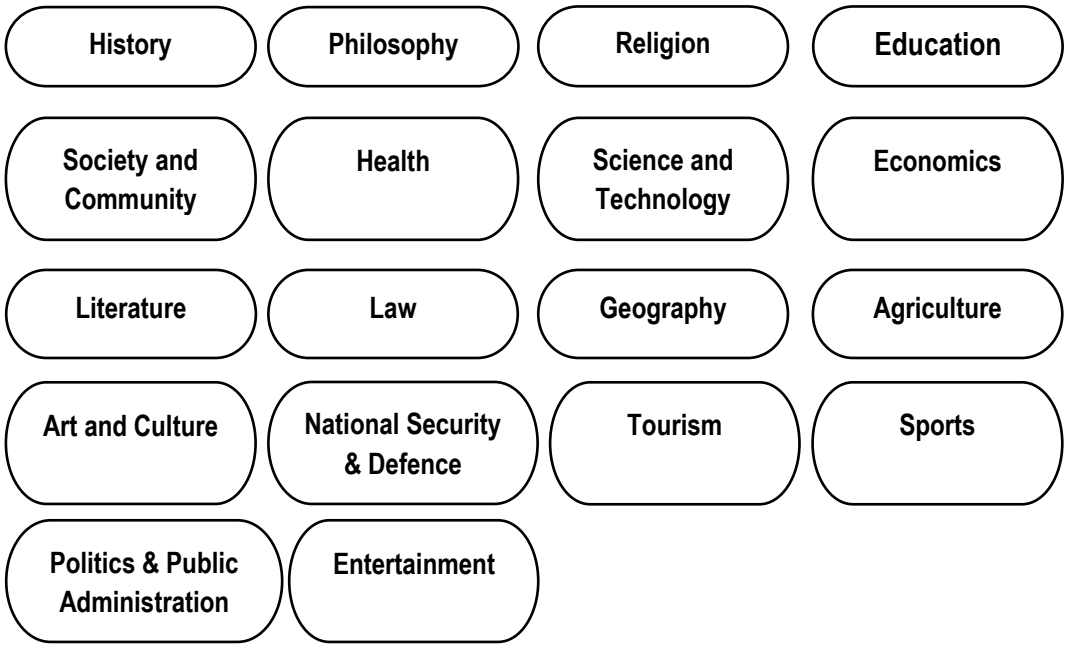

Fig. 1: Major domains of text samples of the Bangla web corpus

Each domain contains several sub-domains. In the project plan it was collectively decided that at least 1000 sentences from each sub-domain should be collected to constitute 5000 sentences for each main domain. Although there are several sub-domains under each main domain, only five sub-domains of each main domain are mentioned below (Table 1).

Main Domain	Sub-domains	Sentence
Agriculture	Agricultural economics/ Agricultural marketing/ Agricultural research/ Agricultural policy/ Crop production	5,000
Art and Culture	Classical performing arts/ Crafts and handicrafts/ Fine arts/ Cuisines/ Costumes	5,000
Economy	Employment/ Industries/ E-commerce/ Labour Economics/ Public finance	5,000

Education	Elementary education/ Secondary education/ Higher education/ Adult education/ Career guidelines	5,000
Entertainment	Film script/ Film reviews/ Media news/ Media personalities/ Film personalities	5,000
Geography	Ancient world/ Europe/ Asia/ Africa/ Graphic representation of earth	5,000
Health	Blood heart and circulation/ Bones, joints and muscles/ Brain and nerve/ Digestive system/ Ear nose and throat	5,000
History	Monuments/ Wars/ Civilisations/ Museums/ Archives	5,000
Law	Criminal law/ Cyber law/ Private law/ Religion and law/ International law	5,000
Literature	Fiction/ Essays/ Drama/ Speeches/ Letters	5,000
National Security and Defence	History/ Personalities/ War technology/ Military law/ International relations	5,000
Philosophy	Movements/ Philosophers/ Theories and schools of thought/ Writings/ Scriptures/	5,000
Politics & Public Administration	Constitution/ Justice/ Governance/ Democracy/ Policies	5,000
Religion	Gods/ Religious text/ Mythology/ Spirituality/ Ancient religions	5,000
Science and Technology	Botany and zoology/ Bioscience and life science/ Discoveries and inventions/ Natural science/ Physics	5000
Society and community	Relationship and kinship/ Marriage/ Child learning/ Area planning/ Public structures	5,000
Sports	Sports events/ Indoor and outdoor games/ Sports persons/ Milestone and records/ Traditional games	5,000
Tourism	Ecotourism/ Leisure tourism/ Heritage tourism/ Dark tourism/ Space tourism	5,000
18 domains	90 sub-domains	90,000

Table 1: Domains and sub-domains of the Bangla Web Corpus

5.3 Data Collection

Primarily we have used two types of basic web sources for data collection. That means two major types of data collection are used to generate this monolingual web corpus: data from structured texts and data from non-structured texts.

5.3.1 Data from Structured Texts

First, we have tried to collect structured texts from web sites of some well known magazines, news papers, and e-books. Data collection from these sources are crucial tasks as it requires high level of persistence in compilation of data in a consistently accurate manner. Moreover, the whole process involves selection of domain specific texts, crawling through digital texts, removal of source code, copying of text in doc files, and text normalization. Furthermore, recurrent maintenance of personal contacts with various newspaper editors and publishers is

also an important task that comes under this method. Finally, the issue of copyright has to be taken care of so that the work of data collection as well as subsequent use of the corpus are no jeopardized. In this context, it may be informed that to avoid any kind of dispute of copy right it is always sensible not to go beyond the limit of 90 words or 1/3 of a whole text of a single piece is work. That means one can extract 90 words or 1/3 of a text for compilation of a web corpus without violating the copy right of the text producers.

5.3.2 Data from Non-Structured Texts

The web offers enormous amount of non-structured texts for corpus development. Such texts are available from wide range of topics, subject domains, and text varieties with unbound limit for data accumulation. It is therefore, is a tough task to restrain oneself in the task of data collection from un-structured web-sources. For our BWC, we have restrained ourselves in data collection mainly from the following sources: emails, web pages, home pages, news portals, and blogs. In a very careful manner, first we had to analyse and attest the relevance of data of a particular site to the basic structure and content of our web corpus, and when it was ascertained, we had culled necessary data and store these in domain-specific files of the corpus. In this case, however, the issue of copyright is not punching us hard, as the data is freely available for general research and development purposes. The entire process of data collection from the web sources is presented through a flow chart in the following manner for better comprehension (Fig. 2).

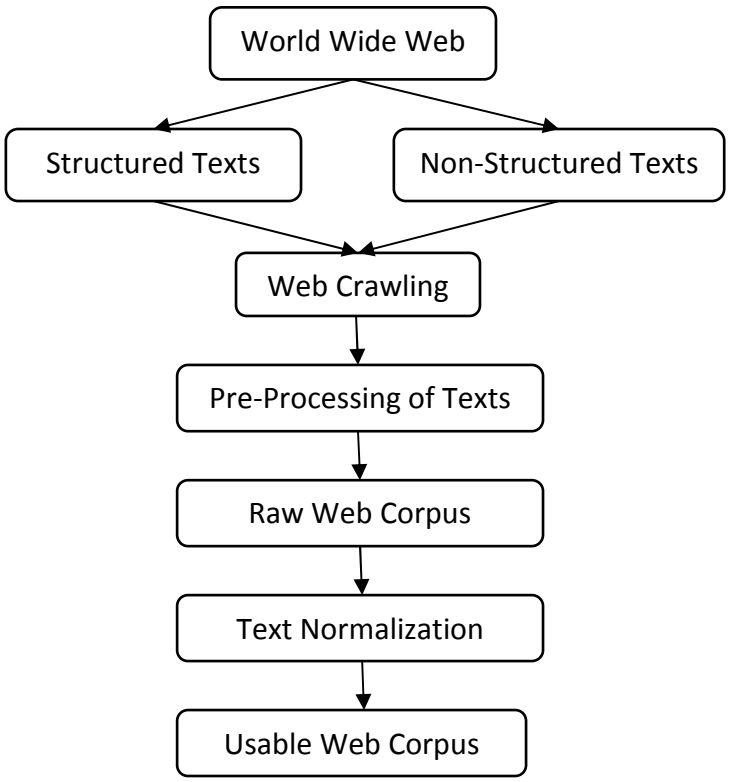

Fig. 2: Stages involved in Bangla Web Corpus (BWC) compilation

6. Metadata Information

Each type of textual data captured and captivated in the web corpus is provided with detailed metadata information for future reference and utilization of the same in the activities of text verification, content classification, text categorization, corpus validation and information retrieval. In this context, it should clearly be noted that due to variation of source, text type, and other factors relating to text generation, metadata information is bound to vary from text to text (Dash 2011). For example, if text data is collected from magazines, then information of volume, number, and year of the source (i.e., magazine) are bound to vary and such variations are mandatory to be recorded in the metadata panel of the corpus. On the other hand, if text data is procured from a book, then the name of the author(s), publisher, main subject area as well as the year of publication are to be furnished in the metadata of the file. Similarly, it is necessary to provide the web address, site name, URL, year, etc. if text data is collected from a web source. On the other hand, if data is collected from a newspaper, then it is mandatory to supply name of the newspaper, place of publication, broad area of the text, date of publication, etc. in the matadata profile. The following four diagrams (Fig. 3, Fig. 4, Fig. 5, Fig. 6) show how metadata information has been furnished with the four major text types of the BWC.

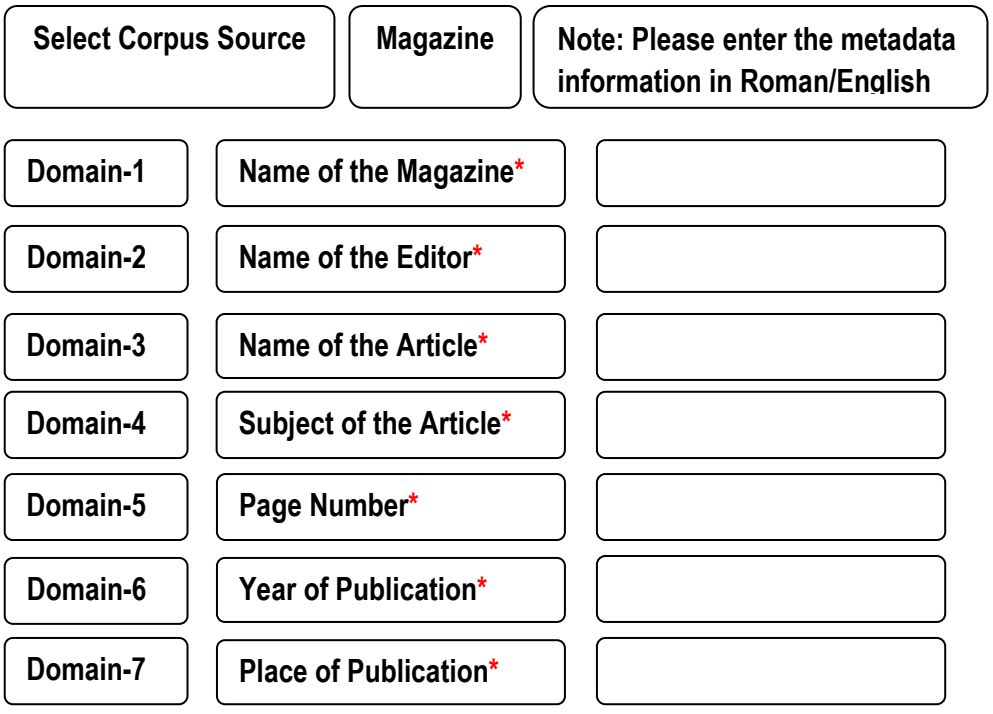

Fig. 3: Metadata information for the texts taken from magazines

Select Corpus Source	Book	Note: Please enter the metadata information in Roman/English
Domain-1	Name of the Book*	
Domain-2	Name of the Editor/Author*	
Domain-3	Name of the Chapter/Article*	
Domain-4	Page Number*	
Domain-5	Name of the Publisher*	
Domain-6	Year of Publication*	
Domain-7	Place of Publication*	

Fig. 4: Metadata information for the texts taken from books

Select Corpus Source	Newspaper	Note: Please enter the metadata information in Roman/English
Domain-1	Name of the Newspaper*	
Domain-2	Name of the Supplement*	
Domain-3	Name of the Article*	
Domain-4	Subject of the Article*	
Domain-5	Page Number*	
Domain-6	Name of the Author*	
Domain-7	Date of Issue (DD-MM-YYYY)*	
Domain-8	Place of Issue*	

Fig. 5: Metadata information for the texts taken from newspapers

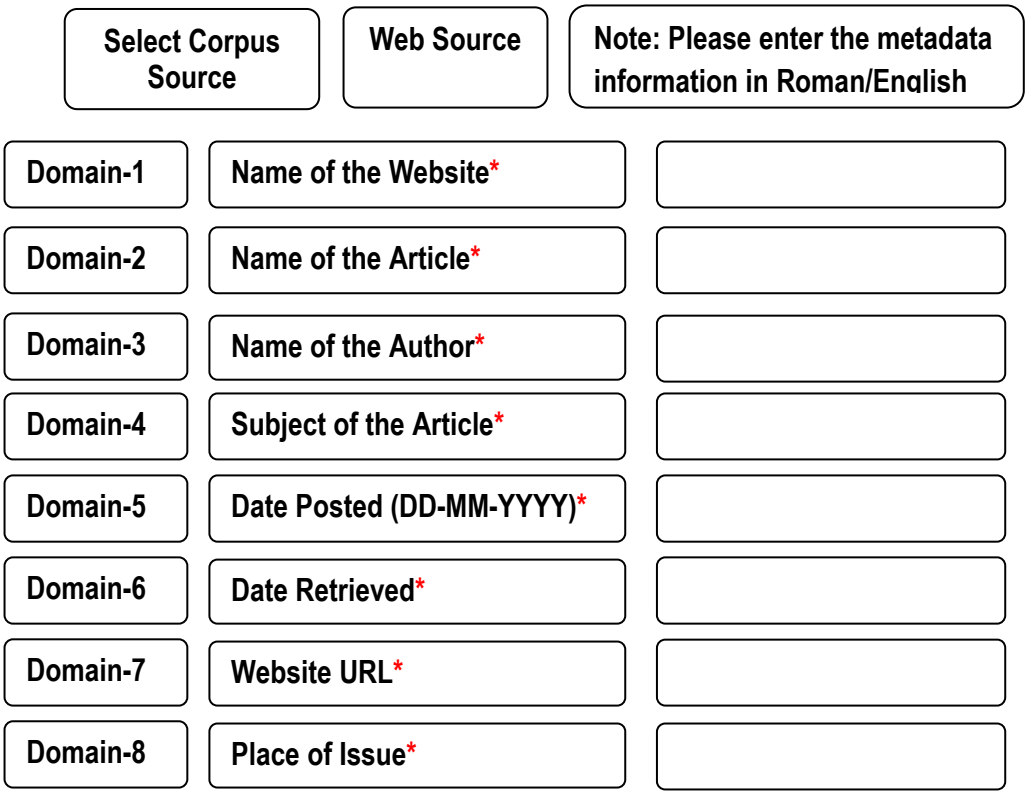

Fig. 6: Metadata information for the texts taken from web sites

6.1 Computerizing the Data

After the web text data collection work is over, appropriate and adequate preparation is made for entering those text data in an electronic format in computer. Although initial planning was made for preserving the data in document (.doc) format, eventually it was found that it is more advantageous to store the text data in Note Pad in UTF8 format so that subsequent text access and processing of the database in various manners and formats are not troublesome. However, the most laborious part of the game is the process of extraction, manipulation, and storage of data in Unicode font format. Also there are problems relating to selection and retrieval of data from various web sources as well as normalisation of the digital texts to make them fit for future usage. Since the process of collection of text data from digital sources is practically different from the process applied for collecting text data from printed sources, one has to be quite innovative in the task of capturing text sources and manoeuvring the text loads in a successful manner. For our purpose, the following two basic strategies were successfully adopted:

(a) Use of a tool called 'Paragraph Splitter' and,
(b) Use of a tool called 'Text Normaliser'.

The first tool helped us capture a text from the web source, store the same in a notepad file, and break the text into manageable paragraphs. The second tool helped us preen the text in a predefined order to remove non-textual elements and materials (e.g., images, tables, diagrams, etc.) to give an acceptable shape to the text. Both the tools are excessively used in sequential order to normalize the text materials in the BWC for subsequent applications. After completion of both the processes the entire raw corpus is stored as a text file in a separate database. It is also uploaded in the central server located at the *Special Centre for Sanskrit Studies, Jawaharlal Nehru University*, New Delhi, India for global access.

6.2 Validation of Web Corpus

The process of validation of the BWC is another crucial phase of corpus management as the utility of a corpus largely depends on certification that the data stored in it is authentic, valid, and true to the language for which its stands for. The process of validation starts just after the completion of the process of corpus compilation and text normalization. The sequential works of corpus generation, normalization, and validation can also be carried out in parallel fashion if a large team is involved in the work through parallel distribution of specific tasks assigned to specialised group members. The underlined argument is — it is the duty of the corpus developers to certify and attest that the web corpus is validated and authenticated for all kinds of application in all possible spheres of linguistics and language technology as well as in other domains of human knowledge.

Alternatively, if experts of the language concerned, are not present in the corpus building team, it is better to hire experts of the languages who have adequate linguistic knowledge to certify authenticity and validity of the text captured in the corpus. In our case, the corpus collectors themselves have validated the texts through the 'cross validation' process monitored and regulated by linguists, senior linguists, and chief investigator. Even then it is always desirable that some external experts should examine and certify the corpus for access so that the cloud of scepticism of biasness is evaporated from the minds of the end users. In case of the BWC, the corpus is now in the process of being available to the experts for further verification, validation and authentication.

7. Problems in Generation of Web Corpus

It is desirable to refer to the hurdles and reflect on the problems that we have come across in the process of web corpus development. In fact, proper reference to these obstacles will not only focus on the complexities involved in the project, but also make the new generation of corpus developers aware about the quicksand under the cyber surface of corpus generation. Here we refer to such problems. In practicality, we had to face two types of problems, as noted below:

(a) Technical problems, and
(b) Linguistic problems.

7.1 Technical Problems

Technical problems are mainly related to non-availability of NLP-trained skilled man power, lack of operation-friendly system interface, lack of data managing customised tool, collection of data from web sources, storing of data in computer, processing of data, and copyright, etc. Some of the problems are addressed below with relevant data and information.

(a) Problem of Data Availability

Getting specific Bangla text data for certain sub-domains was a real big challenge for us. It was difficult to extract data from domains like national security and defence, forensic science, society and community, ethnology, science and technology, which have several sub-domains like war technology, landscape and architecture, palaeontology, palaezoology, genome technology — to mention a few. Either there was not sufficient amount of textual data in the web or the data was encrypted in such a manner that it was not retrievable.

(b) Problem in Download and Storage

While downloading text data on the server we have come across some technical problems. In most cases, the text data is not compatible to the encoding architecture of the Unicode. That means, most of the Bangla text data that are available in the web, are either presented as .pdf text (as noted in case of some Bangla newspapers), or the text is composed in Indian Standard Code for Information Interchange (ISCII) or some other font formats, which are not compatible to Unicode. Such problems have often hampered out work of web corpus generation. At some extreme situations, the downloaded texts are discarded as garbage because no conversion algorithm was able to render these texts into Unicode compatible texts.

(c) Problem of Copyright

Due to copyright constraints it was not possible to collect the entire text database from the websites. Since we had to maintain the rules and norms of copyright of electronic texts, we had to cull text samples in a limited size (one-third of a text or 90 words from a paragraph) to meet our needs. In case of those free texts, where there is no question of copyright, we have taken full liberty to download textual data as much as we could to meet our target.

7.2 Linguistic problems

The linguistic problems, on the other hand, are mainly related to orthography (i.e., spelling), grammar, lexical form, affixation, dialectal variation, punctuation, discourse, domain overlap, usage, etc. Some of the linguistic problems are discussed below:

(a) Spelling Errors

The web corpus developers have noted several spelling errors in the text samples selected for the corpus. These are not spelling variations, these are actual spelling errors. These are

corrected manually by the corpus developers so that the correct forms of words are stored in the database. For elucidation, let us consider some of the errors given below as examples:

Wrong form : নিশাত হাত বাড়িয়ে বৃষ্টির ফোটা স্পর্শ করলো ।
Roman : niśāt hāt bāṛiye bṛṣṭir phoṭā sparśa karlo.
Correct form : নিশাত হাত বাড়িয়ে বৃষ্টির ফোঁটা স্পর্শ করলো ।
Roman : niśāt hāt bāṛiye bṛṣṭir phõṭā sparśa karlo.
Meaning : By stretching his hand Nishat touched the raindrop.

(b) Syntactic Errors

Syntactic errors are commonly found when the grammatical concord between the subject and predicate is lost. The responsibility of a corpus compiler is to correct such error. For example,

Wrong form : তিনি সেখানে বসে পড়ল ।
Roman : tini sekhāne base paṛla.
Correct form : তিনি সেখানে বসে পড়লেন ।
Roman : tini sekhāne base parlen.
English : He (non-Hon) sat down there.

(c) Use of Informal Words

It is noted that in some cases, an informal or colloquial form of a word is used in the standard or formal version of a text. This is normally known as 'gurucaṇḍālī doṣ' (fallacy of random cross-formal lexical mix), as the following example shows:

Wrong form : কিছু কথা কইতে চাই ।
Roman : kichu kathā kaite cāi.
Correct form : কিছু কথা বলতে চাই ।
Roman : kichu kathā balte cāi.
English : I want to say something.

Wrong form : সে ছুটিতে গেরামে গেছিল ।
Roman : se chuṭite gerāme gechila.
Correct form : সে ছুটিতে গ্রামে গেছিল ।
Roman : se chuṭite grāme gechila.
English : He went to village during vacation.

In such cases, we have tried to collect text data from those : web sources where standard and formal Bangla texts are available. It is to be noted that if text data is collected from the websites of Bangladesh, then we may came across a lot of terms and words that are found in Bangla used in Bangladesh but not available in Bangla used in West Bengal, India (e.g., সাদি (sādi) : বিয়ে (biye) "marriage", পানি (pāni) : জল (jal) "water", মরিচ (maric) : লংকা (lankā) "chilli", রসুই (rasui) : রান্না (rānnā) "cooking", ফুফু (phuphu) : পিসি (pisi) "father's sister", দাওয়াত (dāoyāt) : নিমন্ত্রণ (nimantraṇ) "invitation", etc.). This particular issue is not addressed here due to its wide controversial identity.

(d) Punctuation Errors

In some texts, punctuation markers like full stop, comma, dash, etc. are not used properly. As a result, either two or more separate sentences are joined together without any overt connectors, else, one single sentence is broken into two or more separate sentences without any reason or logic. Similar misuse is also noted in case of other punctuation marks, mostly for hyphen, comma, semicolon, colon. etc. In most cases, the corpus compiler has to put appropriate punctuation marks at appropriate places after reading out the text in the corpus.

(e) Problem in Maintaining Discourse

Maintaining discourse continuation is a major problem due to copyright issues in corpus development. As per copyright rules and regulations we are supposed to extract only 90 words or 1/3rd of the whole text at a time. Due to this reason the logical link between two or more paragraph sequences is often snapped and as a result of this it has become difficult to establish and maintain discourse continuation link in a piece of text. Since this is a logistic problem where a corpus designer has hardly any role to play, it makes the whole process of maintaining discursive relationship across text sequences a real linguistic challenge.

(f) Problem of Overlapping Domains

As we are dealing with various genres regarding monolingual corpus including 18 different domains (which has many sub-domains too), the problem of text overlapping across sub-domains is a problem of novel type. It is better to call it a 'text identity problem'. In our case, for instance, there are text samples, which at the same time, may belong to both religion and philosophy, or tourism and history, or nature and geography, or music and culture, etc. That means, text belonging to a head domain may often overlap across several sub-domains of the head domain. For instance, a text of travelogue may belong to literature as well as travelogue. Similarly a text on cinema may belong to film and entertainment. The same situation arises when we deal with the sub-domains of classical performing arts under 'art and culture', which overlaps with the sub-domain of performing arts under 'entertainment'. Such problems may be solved through detailed analysis of texts for recategorization.

8. Conclusion

In this paper, we have tried to present a short description about the process that we have used to generate a web corpus in Bangla — a type of corpus of its own — which has never been attempted before in Bangla or other Indian languages. This paper also discusses the strategies we have deployed as well as the challenges that we have faced during this process. In the course of the work the following route-map is successfully followed for collecting the web data: crawling the webs — harvesting the sites — collecting digital texts — storing the texts — and generating the web corpus. It is a new route, which is full of meanders, hardly known, and sparsely traversed.

The value of this web corpus will increase over the years and it will be regarded as one of the most useful resources for multiple linguistic research and investigations. We believe that this BWC will open up many new avenues of studies in language technology, communication, and

linguistics. If we succeed to annotate this corpus, if will far more useful in many domains of human knowledge eventually leading to development of various linguistic tools and resources for Bangla.

The World Wide Web (WWW), which is visualised here as a useful linguistic resource, in itself, is a unique linguistic world full of surprising linguistic data and information. In fact, it is the largest store of texts in existence, freely-available, covering a wider range of domains, and constantly added to and updated by one and all (Renouf 2003: 40). This huge collection of text, if properly processed and annotated, can be highly useful in linguistic and non-linguistic studies, cross linguistic comparisons, language technology, and all other domains of descriptive, theoretical and applied linguistics.

Finally, we visualize that in the long run, along the side of corpora generated from printed texts, corpora produced from web texts may be equally used in natural language processing, linguistic resource development, cross-lingual communication, globalization of linguistic profiles and language resources, digital lexical database, computational lexicography, language planning and E-governance.

References

Atkins, Sue, Jereme Clear and Nicholas Ostler (1992) Corpus design criteria. *Literary and Linguistic Computing*. 7(1): 1-16.

Baroni, Marco (2005) Large crawls of the web for linguistic purposes. Workshop paper presented at *Corpus Linguistics 2005*, Birmingham, UK.

Bergh, Gunnar, Aimo Seppänen, and Joe Trotta (1998) Language corpora and the internet: a joint linguistic resource. In: Antoinette Renouf (ed.) *Explorations in Corpus Linguistics*, pp. 41-54. Amsterdam/Atlanta: Rodopi.

Brekke, Magnar (2000) From the BNC toward the cyber corpus: a quantum leap into chaos? In: J.M. Kirk (Ed.) *Corpora Galore: Analyses and techniques in describing English*. Proceedings of the 19th International Conference on English Language Research on Computerised Corpora (Language and Computers 30), pp. 227-247. Amsterdam and Atlanta: Rodopi.

Cavaglia, Gabriela and Adam Kilgarriff (2001) Corpora from the web. *Information Technology Research Institute Technical Report Series* (ITRI-01-06): ITRI, University of Brighton, UK.

Chang, Jing-Shin (2005) Domain specific word extraction from hierarchical web documents: a first step towards building lexicon trees from web corpora. *Proceedings of the 4th SIGHAN Workshop on Chinese Language Learning,* pp. 64-71, as a part of IJCNLP-05 (International Joint Conference on Natural Language Processing), Jeju Island, Korea, October 14-15, 2005.

Dash, Niladri Sekhar (2004) Language corpora: present Indian needs. *Proceedings of the SCALLA 2004 Working Conference*: Crossing Digital Divides shaping technologies to meet human needs, Kathmandu, Nepal, 5-7 January 2004. http://www.elda.fr/proj/scalla.html.

Dash, Niladri Sekhar (2005) *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.

Dash, Niladri Sekhar (2007) Indian scenario in language corpus generation. In: Dash, Niladri Sekhar Dash, Probal Dasgupta, and Pabitra Sarkar (Eds.) *Rainbow of Linguistics: Vol. I.,* pp. 129-162. Kolkata: T. Media Publication.

Dash, Niladri Sekhar (2008) *Corpus Linguistics: An Introduction*. New Delhi: Pearson Education-Longman.

Dash, Niladri Sekhar (2009) *Language Corpora: Past, Present and Future*. New Delhi: Mittal Publications.

Dash, Niladri Sekhar (2011) Extratextual (documentative) annotation in written text corpora. *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)*, pp. 168-176, Anna University, Chennai, India, 16-19 December 2011.

Dash, Niladri Sekhar (2012) From KCIE to LDC-IL: some milestones in NLP journey in Indian multilingual panorama. *Indian Linguistics*. 73(1-4): 129-146.

Ekbal, Asif and Sivaji Bandyopadhyay (2008) Web based Bengali news corpus for lexicon development and POS tagging. *POLIBITS*. 37(1): 20-29.

Fletcher, William (2004) Making the web more useful as a source for linguistic corpora. In: U. Connor and T. Upton (Eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*, pp. 191-205. Amsterdam: Rodopi.

Francis, W. Nelson and Henry Kucera (1964) *Manual of information to accompany A standard Corpus of present-day edited American English*. Dept. of Linguistics, Brown University, Providence, R.I.: USA.

Ghani, Rayid, Rosie Jones and Dunja Mladenic (2003) Building minority language corpora by learning to generate Web search queries. *Knowledge and Information Systems*. 7(1): 56-83.

Ide, Nancy, Randi Reppen, and K. Suderman (2002) The American National Corpus: more than the web can provide. *Proceedings of the 3rd Language Resources and Evaluation Conference (LREC), Canary Islands*. Paris: ELRA.

Kehoe, Andrew (2006) Diachronic linguistic analysis on the web with WebCorp. In: Antoinette Renouf and Andrew Kehoe (Eds.) *The Changing Face of Corpus Linguistics*, pp. 297-307. Amsterdam/New York: Rodopi.

Kehoe, Andrew and Antoinette Renouf (2002) WebCorp: applying the web to linguistics and linguistics to the web. *World Wide Web 2002 Conference,* Honolulu, Hawaii, 7-11 May 2002, http://www2002.org/CDROM/poster/67/

Kehoe, Andrew and M. Gee (2007) New corpora from the web: making web text more 'text-like'. In: P. Pahta, I. Taavitsainen, T. Nevalainen and J. Tyrkkö (eds.) *Studies in Variation, Contacts and Change in English Volume 2: Towards Multimedia in Corpus Studies,* University of Helsinki e-journal.

Kehoe, Andrew and M. Gee (2009) Weaving web data into a diachronic corpus patchwork. In: Antoinette Renouf and Andrew Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments*, pp. 255-279. Amsterdam: Rodopi.

Kilgarriff, Adam (2001) Web as corpus. *Proceedings of the Corpus Linguistics 2001 Conference*, Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew; and Khoja, S. (Eds.) pp. 342-344, UCREL, 2001.

Kucera, Henry and W. Nelson Francis (1967) *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Lawrence, Steve and C. Lee Giles (1998) Searching the World Wide Web. *Science,* 280: 98-100.

Meyer, Charles, Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and Stephanie Moses (2003) The World Wide Web as linguistic corpus. In: P. Leistyna and C.F. Meyer

(Eds.) *Corpus analysis. Language structure and language use* (Language and Computers 46), pp. 241-254. Amsterdam and New York: Rodopi.

Pala, Kiran and Suryakanth V. Ganagashetty (2012) Challenges and opportunities in automatically building bilingual lexicon from web corpus. *Interdisciplinary Journal of Linguistics.* 5(1-2): 169-184.

Renouf, Antoinette (2003) WebCorp: providing a renewable data source for corpus linguists. In: S. Granger and S. Petch-Tyson (eds.) *Extending the scope of corpus-based research: new applications, new challenges*, pp. 39-58. Amsterdam and New York: Rodopi.

Renouf, Antoinette (2003) WebCorp: providing a renewable data source for corpus linguists. In: S. Granger and S. Petch-Tyson (eds.) *Extending the scope of corpus-based research. New applications, new challenges* (Language and Computers 48), pp. 39-58. Amsterdam and New York: Rodopi.

Renouf, Antoinette and Andrew Kehoe (2013) Filling the gaps: using the WebCorp linguist's search engine to supplement existing text resources**.** *International Journal of Corpus Linguistics*, 18(2): 167-198.

Renouf, Antoinette, Andrew Kehoe, and D. Mezquiriz (2004) The accidental corpus: issues involved in extracting linguistic information from the web. In: Karim Aijmer and Bengt Altenberg (eds.) *Advances in Corpus Linguistics*, pp. 403-419. Amsterdam: Rodopi.

Resnik, Philip and Noah A. Smith (2003) The web as a parallel corpus. *Computational Linguistics.* 29(3): 349-380.

Sarma, Shikhar, Himadri Bharali, Ambeswar Gogoi, Ratul Deka, and Anup Barman (2012) A structured approach for building Assamese corpus: insight, application and challenges. Presented in the *24th International Conference on Computational Linguistics (COLING 2012),* 8-15 December 2012, Mumbai, India.

Sharoff, Serge (2006) Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics.* 11(4): 435-462.