Corpus Linguistics : A General Introduction

Niladri Sekhar Dash

1. Introduction

Corpus Linguistics is a multidimensional area. It is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication, and comprehension. The introduction of corpus in language study and application has incorporated a new dimension to linguistics.

In principle, *Corpus Linguistics* is an approach that aims at investigating language and all its properties by analysing large collections of text samples. This approach has been used in a number of research areas for ages: from descriptive study of a language, to language education, to lexicography, etc. It broadly refers to exhaustive analysis of any substantial amount of authentic, spoken and/or written text samples. In general, it covers large amount of machine-readable data of actual language use that includes the collections of literary and non-literary text samples to reflect on both the synchronic and diachronic aspects of a language.

The uniqueness *corpus linguistics* lies in its way of using modern computer technology in collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in application of these in all kinds language-related research and development activities.

Electronic (digital) language corpus is a new thing. It has a history of nearly half a century. Therefore, we are yet to come to a common consensus as to what counts as corpus, and how it should be designed, developed, classified, processed and utilised.

The basic philosophy behind corpus linguistics has two wings: (a) we have a cognitive drive to know how people use language in their daily communication activities, and (b) if it is possible to build up intelligent systems that can efficiently interact with human beings. With this motivation both computer scientists and linguists have come together to develop language corpus that can be used for designing intelligent systems (e.g., machine translation system, language processing system, speech understanding system, text analysis and understanding system, computer aided instruction system, etc.) for the benefit of the language community at large.

All branches of linguistics and language technology can benefit from insights obtained from analysis of corpora. Thus, description and analysis of linguistic properties collected from a corpus becomes of paramount importance in all many areas of human knowledge and application.

2. What is Corpus?

The term ***corpus*** is derived from Latin ***corpus*** "body". At present it means representative collection of texts of a given language, dialect or other subset of a language to be used for linguistic analysis. In finer definition, it refers to (a) (loosely) any body of text; (b) (most commonly) a body of machine-readable text; and (c) (more strictly) a finite collection of machine-readable texts sampled to be representative of a language or variety (McEnery and Wilson 1996: 218).

Corpus contains a large collection of representative samples of texts covering different varieties of language used in various domains of linguistic interactions. Theoretically, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. It is compatible to computer, operational in research and application, representative of the source language, processable by man and machine, unlimited in data, and systematic in formation and representation (Dash 2005: 35).

3. Salient Features of Corpus

- **Quantity:** It should be big in size containing large amount of data either in spoken or written form. Size is virtually the sum of its components, which constitute its body.

- **Quality (=** authenticity). All texts should be obtained from actual examples of speech and writing. The role of a linguist is very important here. He has to verify if language data is collected from ordinary communication, and not from experimental conditions or artificial circumstances.

- **Representation:** It should include samples from a wide range of texts. It should be balanced to all areas of language use to represent maximum linguistic diversities, as future analysis devised on it needs verification and authentication of information from the corpus representing a language.

- **Simplicity:** It should contain plain texts in simple format. This means that we expect an unbroken string of characters (or words) without any additional linguistic information marked-up within texts. A simple plain text is opposed to any kind of annotation with various types of linguistic and non-linguistic information.

- **Equality:** Samples used in corpus should be of even size. However, this is a controversial issue and will not be adopted everywhere. Sampling model may change considerably to make a corpus more representative and multi-dimensional.

- **Retreavability:** Data, information, examples, and references should be easily retrievable from corpus by the end-users. This pays attention to preserving techniques of language data in electronic format in computer. The present technology makes it possible to generate corpus in PC and preserve it in such way that we can easily retrieve data as and when required.

- **Verifiability:** Corpus should be open to any kind of empirical verification. We can use data form corpus for any kind of verification. This puts corpus linguistics steps ahead of intuitive approach to language study.

- **Augmentation:** It should be increased regularly. This will put corpus 'at par' to register linguistic changes occurring in a language in course of time. Over time, by addition of new linguistic data, a corpus achieves historical dimension for diachronic studies, and for displaying linguistic cues to arrest changes in life and society.

- **Documentation:** Full information of components should be kept separate from the text itself. It is always better to keep documentation information separate from the text, and include only a minimal header containing reference to documentation. In case of corpus management, this allows effective separation of plain texts from annotation with only a small amount of programming effort.

4. The TDIL Corpora of Indian Languages

The generation of corpus for the Indian languages began in 1991, when the *Department of Electronics, Govt. of India* initiated a project (*Technology Development for Indian Languages*) to develop machine-readable corpus of texts in all the major Indian languages. Also emphasis was laid on development of software for language processing (e.g., POS tagging, text encoding, frequency counting, spelling checking, morphological processing, etc.) as well as for designing of systems for machine translation from English to Indian languages.

Indian Institute of Technology, Delhi was given the task for developing corpus of Indian English, Hindi, and Punjabi; *Central Institute of Indian Languages*, Mysore was assigned for corpus of Tamil, Telugu, Kannada, and Malayalam; *Deccan College*, Pune developed corpus of Marathi and Gujarati; *Indian Institute of Applied Language Sciences*, Bhubaneswar developed corpus of Oriya, Bengali and Assamese; *Sampurnananda Sanskrit University*, Varanasi was entrusted with the development of corpus of Sanskrit; and *Aligarh Muslim University*, Aligarh was assigned the task for developing corpus of Urdu, Sindhi, and Kashmiri; *Indian Institute of Technology*, Kanpur took responsibility for designing systems and software for language processing and machine translation, while CIIL, Mysore took responsibility to archive the entire corpus database of all Indian languages for future utilization.

After the completion of the TDIL project in 1995, works on corpus development and processing stopped for some reasons beyond the knowledge of the present author. However, the *MICT, Govt. of India* has revived the whole enterprise (http:\\www.mit.gov.in) with new enthusiasm and vision. The realisation of this enterprise is clearly manifested in formation of LDC-IL, although my personal view is that it should be **NAIL** (National Archive for the Indian Languages), rather than the LDC-IL (Dash 2003).

5. Conceptual Classification of Corpora

Since electronic corpus is a new thing, we are yet to reach to a common consensus to what counts as a corpus, and how it should be classified. The classification scheme I propose here goes as far as it is prudent at the present moment. It offers a reasonable way to classify corpora, with clearly delimited categories wherever possible. Different criteria for classification are applied to corpora, sub-corpora, and their related components. Linguistic criteria may be external and internal. External criteria are largely mapped onto corpora from text typology concerned with participants, occasion, social setting, communicative function of language, etc. Internal criteria are concerned with recurrence of language patterns within the pieces of language. Taking all these issues under consideration I classify corpora in a broad scheme in the following manner: Genre of text, Nature of data, Type of text, Purpose of design, and Nature of application.

5.1 Genre of Text

- **Written Corpus:** A written corpus (e.g., TDIL Corpus) by virtue of its genre contains only language data collected from various written, printed, published and electronic sources.

- **Speech Corpus:** A speech corpus (e.g., *Wellington Corpus of Spoken New Zealand English*) contains all formal and informal discussions, debates, previously made talks, impromptu analysis, casual and normal talks, dialogues, monologues, various types of conversation, on line dictations, instant public addressing, etc. There is no scope of media involvement in such texts.

- **Spoken Corpus:** Spoken corpus (e.g., *London-Lund Corpus of Spoken English*), a technical extension of speech corpus, contains texts of spoken language. In such corpus, speech is represented in written form without change except transcription. It is annotated using a form of phonetic transcription.

5.2. Nature of Data

- **General Corpus:** General corpus (e.g., *British National Corpus*) comprises general texts belonging to different disciplines, genres, subject fields, and registers. Considering the nature of its form and utility, it is finite in number of text collection. That means, number of text types and number of words and sentences in it are limited. It has an opportunity to grow over time, and to append new data with availability of new texts. It is very large in size, rich in variety, wide and representation, and vast in utilisation scope.

- **Special Corpus:** Special corpus (e.g., *CHILDES Database*) is designed from texts sampled in general corpus for specific variety of language, dialect and subject with emphasis on certain properties of the topic under investigation. It varies in size and composition according to purpose. It does not contribute to the description of a language because it contains a high proportion of unusual features. Its origin is not reliable as it records the data from people not behaving normally. Special corpus is not balanced (except within the scope of its given purpose) and, if used for other purposes, gives distorted and 'skewed' view of language segments. It is different in principle, since it features one or other variety of normal, authentic language. Corpus of language of children, non-native speakers, users of dialects, and special areas of communication (e.g., auction, medical talks, gambling, court proceeding, etc.) are designated as special corpus because of their non-representative nature of the language involved. Its main advantage is that texts are selected in such a way that the phenomena one is looking for occur more frequently in it than in balanced corpus. A corpus that is enriched in such a way is smaller than a balanced corpus providing same type of data (Sinclair 1996b).

- **Sublanguage corpus:** It consists of only one text variety of a particular language. It is at the other end of the linguistic spectrum of a Reference corpus. The homogeneity of its structure and specialised lexicon allows the quantity of data to be small to demonstrate typically good and closure properties.

- **Sample corpus:** Sample corpus (e.g., *Zurich Corpus of English Newspapers*) is one of the categories of special corpus, which is made with samples containing finite collection of texts chosen with great care and studied in detail. Once a sample corpus is developed it is not added to or changed in any way (Sinclair 1991: 24) because any kind of change will imbalance its constitution and distort research requirement. Samples are small in number in relation to texts, and of constant size. Therefore, they do not qualify as texts.

- **Literary corpus:** A special category of sample corpus is literary corpus, of which there are many kinds. Classification criteria considered for generation of such corpus include author, genre (e.g., odes, short stories, fictions, etc.), period (e.g., 15th century, 18th century, etc.), group (e.g., Romantic poets, Augustan prose writers, Victorian novelists, etc.), theme (e.g., revolutionary writings, family narration, industrialisation, etc.) and other issues as valued parameters.

- **Monitor corpus:** Monitor corpus (e.g., *Bank of English*) is a growing, non-finite collection of texts with scope for constant augmentation of data reflecting changes in language. Constant growth of corpus reflects change in language, leaving untouched the relative weight of its components as defined by parameters. The same composition schema is followed year by year. The basis of monitor corpus is of reference to texts spoken or written in one single year (Sinclair 1991: 21). From monitor corpus we find new words, track variation in usage, observe change in meaning, establish long-term norm of frequency distribution, and derive wide range of lexical information. Over time the balance of components of a monitor corpus changes because new sources of data become available and some new procedures enable scarce material to become plentiful. The rate of flow is adjusted from time to time.

5.3 Type of Text

- **Monolingual corpus:** It (e.g., *ISI Bengali Corpus*) contains representative texts of a single language representing its use in a particular period or in multiple periods. It contains both written and spoken text samples so long their cohabitation and relational interface does not hamper proposed work of the investigators.

- **Bilingual corpus:** Bilingual corpus (e.g., *TDIL Bengali-Oriya Corpus*) is formed when corpora of two related or non-related languages are put into one frame. If these languages are genetically or typologically related they become parallel corpus (discussed below) where texts are aligned

following some predefined parameters. Size, content, and field may vary from corpus to corpus, which is not permitted in case of parallel corpus.

- **Multilingual corpus:** Multilingual corpus (e.g., *Crater Corpus*) contains representative collections from more than two languages. Generally, here as well as in bilingual corpus, similar text categories and identical sampling procedures are followed although texts belong to different languages.

5.4 Purpose of Design

- **Un-annotated corpus:** It (e.g., TDIL Corpus) represents a simple raw state of plain texts without additional linguistic or non-linguistic information. It is of considerable use in language study, but utility of corpus is considerably increased by annotation.

- **Annotated corpus:** It (e.g., *British National Corpus)* contains tags and codes inserted from outside by designers to record some extra information (analytical marks, parts-of-speech marks, grammatical category information, etc.) into texts. In contrast to un-annotated corpus, annotated corpus is more suitable for providing relevant information useful in various tasks for language technology including morphological processing, sentence parsing, information retrieval, word sense disambiguation, machine translation, etc.

5.6 Nature of Application

- **Translation Corpora:** Translation corpora generally consist of original texts of source language and their translations taken from target language. These corpora usually keep meaning and function of words and phrases constant across languages, and as a consequence, offer an ideal basis for comparing realisation of particular meanings in two different languages under identical condition. Moreover, they make it possible to discover all cross-linguistic variants, i.e. alternative renderings of particular meanings and concepts. Thus, translation corpora provide more fruitful resources both for cross-linguistic data analysis and rule formulation necessary for translation (Altenberg and Aijmer 2000: 17).

- **Aligned corpus:** It is (e.g., *The Canadian Hansard Corpus*) a kind of bilingual corpus where texts in one language and their translations into other language are aligned, sentence by sentence, phrase by phrase, or even word by word.

- **Parallel corpus:** Parallel corpus (e.g., *Chemnitz German-English Corpus*) contains texts as and translations in each of the languages involved allowing double-checking translation equivalents. Texts in one language and their translations into another are aligned: sentence by sentence, phrase by phrase, or even word by word. Sometimes reciprocate parallel corpora are designed where corpora containing authentic texts and translations in each of the languages are involved.

- **Reference corpus:** It (e.g., *Bank of English*) is made to supply comprehensive information about a language. It is large enough to represent all relevant varieties of language and characteristic vocabulary, so that it can be used for writing grammars, dictionaries, thesauruses and other materials. It is composed on the basis of relevant parameters agreed upon by linguistic community. It includes spoken and written, formal and informal language representing various social and situational registers. It is used as a 'benchmark' for lexicons, for performance of generic tools, and language technology applications. With growing influence of internal criteria, reference corpus is used to measure deviance of special corpus.

- **Comparable corpus:** It is (e.g., *Corpus of European Union*) a collection of 'similar' texts in more than one language or variety. It contains texts in different languages where texts are not same in content, genre, or register. These are used for comparison of different languages. It follows same

composition pattern but there is no agreement on the nature of similarity, as there are few examples of comparable corpora. It is indispensable for comparison in different languages and in generation of bilingual and multilingual lexicons and dictionaries.

- **Opportunistic corpus:** An opportunistic corpus stands for inexpensive collection of electronic texts that can be obtained, converted, and used free or at a very modest price; but is often unfinished and incomplete. Therefore, users are left to fill in blank spots for themselves. Their place is in situations where size and corpus access do not pose a problem. The opportunistic corpus is a virtual corpus in the sense that selection of an actual corpus (from opportunistic corpus) is up to the needs of a particular project. Monitor corpus generally considered as opportunistic corpus.

There can be some other types of specification such as *closed corpus, synchronic corpus, historical corpus, dialect corpus, idiolect corpus,* and *sociolect corpus,* etc. Therefore, the scheme of classification presented here is not absolute and final. It is open for re-categorisation as well as for sub-classification according to different parameters.

6. Text Corpus Generation

There are various issues related with corpus design, development, and management. The issues of corpus development and processing may vary depending on the type of corpus and the purpose of use. Issues related to speech corpus development differ from issues related to text corpus development. Developing a speech corpus involves issues like propose of use, selection of informants, choice of settings, manner of data-sampling, manner of data collection, size of corpus, problem of transcription, type of data encoding, management of data files, editing of input data, processing of texts, analysis of texts, etc.

Developing a written text corpus involves issues like size of corpus, text representation, question of nativity, determination of target users, selection of time-span, selection of documents, collection of text documents (e.g., books, newspapers, magazines etc.), method of data sampling (e.g., sorting of collected materials according to one's need), manner of data input (e.g., random, regular, selective, etc.), corpus sanitation (e.g., error correction omission of foreign words, quotations, dialects etc.), corpus file management, problem of copy-right, etc.

6.1 Size of Corpus

How big a corpus will be? This implies that size is an important issue in corpus generation. It is concerned with total number of words (tokens) and different words (types) to be taken into a corpus. It also involves the decision of how many categories we like keep in the corpus, how many samples of texts we need put into each category, and how many words we shall keep in each sample. Although the question of size affects validity and reliability of a corpus, it is stressed that any corpus, however big, is nothing more than a minuscule sample of all speech and writing varieties produced by the users of a language.

In early days of corpus generation, when computer technology for procuring language data was not much advanced, it was considered that a corpus containing one million words is large enough to represent a language or variety. But by the middle of 1980s, computer technology went through a vast change with unprecedented growth of its storage, processing, and accessing abilities that have been instrumental in changing the concept regarding the size of a corpus.

Now it is believed that the bigger the size of corpus the more it is faithful in representing the language under consideration. With advanced computer technology we can generate corpus of very large size containing hundreds of million of words. For instance, the *Bank of English, British National Corpus, the COBUILD Corpus,* the *Longman/Lancaster Corpus, the International Corpus of English,*

the *American National Corpus, etc.* are indeed very large in size – each one containing more than hundred million words.

6.2 Text Representation

In the long run, however, the question of size becomes irrelevant in the context of representation of text samples. Large corpus does not necessarily represent a language or a variety any better way than a small corpus which is properly balanced with texts of all language varieties. Any large collection of texts is not necessarily a corpus, until it is formed with proper text representation for any kind of generalisation. We can call a corpus 'truly representative' only when the findings obtained from its analysis can be generalised to language as a whole or to a specific part of it. Therefore, rather than focussing on quantity or amount of data in a corpus, it is always better to emphasise on the quality of data or text samples. That means data should be proportionately represented from all possible domains of language use within a corpus. In essence, the overall corpus size and balance need to be set against the diversity of source texts for achieving proper text representation.

Within any text category, the greater is the number of individual text samples, the greater is the reliability of analysis of linguistic variables of a corpus. For instance, the *Brown* and the *LOB Corpus*, as well as the *Survey of English Usage* are considered to be good representatives of English used in America and England because these are carefully designed with systematic representation of modern English texts. A simple comparison of these with the *BNC* (a corpus of 500 million words) will show that the *BNC* is not only larger with regard to the amount of texts, but also far more diversified in structure and in text representation. This easily settles empirically the issue of size and representation of text samples in a corpus.

6.3 Question of Nativity

The question is – whose texts (speech and writing) should be included in a corpus: the texts of the native users or that of the non-native users? The general argument is that if it is a monitor corpus, then texts produced by the native users should get priority over the texts produced by the non-native users. Because, the main aim of a monitor corpus is to represent the language or the variety, which can be considered as an 'ideal form' for all kinds of works in linguistics and language technology.

It has been observed that the citation of 'made-up examples' and the listing of 'ungrammatical' sentences in a monitor corpus have fairly significant effect on the results of linguistic analysis of a corpus. In that case, we find large number of 'mention' rather than 'actual use' of words and phrases within a corpus.

If our main objective for building a corpus is to know the naturally occurring language, in order to see what does occur and what does not, then permitting 'made-up examples of sentences and phrases' will make a corpus less acceptable for proposed purposes. One way of avoiding this and many other potential problems, which may be found in a special corpus, is to apply a criterion for inclusion of texts in corpus that the texts should not be too technical in nature.

In case of special corpus, texts produced by non-native users may be considered, since the aim of a special corpus is to highlight the peculiarities typical to the non-native users. Here the question of representation of texts in a corpus is not related with the language as a whole, but with the language used by a particular class of people who have learnt and used language as their second language.

The basic idea is to have a corpus that includes data from which we can gather information about how a language is commonly used in various mainstream activities of linguistic interaction. When we try to produce references that will provide faithful guidance to word use, spelling variations, syntactic constructions, meanings, usages, etc. most likely, we would like to acquire texts produced by the native users. In principle, the texts written and spoken by the native users will be far more directive, appropriate, and representative for enhancing the ability of language understanding and use for the non-native users. Perhaps, this goes rightly along with the desire of the non-native users, who while learning a second language, aim to achieving the efficiency of a native language user.

However, the question of nativity becomes complicated and case-sensitive when we find that the same language is used by two different speech communities separated by geographical or political

distance (e.g., *British English* and *American English*). In this case, we need to tread on a different track. It is better to recognise and generate separate corpus with separate lexical items and syntactic constructions that are common in, or typical of, the native speakers – especially those which differ from one another (e.g., words and sentences typical to *British English* vs. words and sentences typical to *American English*).

We also need to get into the things that are correct by the 'rules' of grammar and usage of the American English, and perfectly understandable; but just not 'right' in rules of grammar and usage in the British English. This usually betrays the most proficient 'native' speaker of *American English* the opportunity for enlisting their languages in corpus of *British English*.

In this context let us consider about the *Indian English*. When Indian people are exposed to lots of linguistic materials that show the marks of being *non-Indian English* (as Indians are often exposed to many texts of the *British English* and the *American English*), people who want to describe, recognise, understand, and generate *Indian English* will definitely go for the texts produced by native speakers of *Indian English*, which will highlight the linguistic traits typical to *Indian English*, and thus will defy the all pervading influence of the *British English* or the *American English* over the *Indian English*.

6.4 Identification of Target Users

There are no fixed target users for a general corpus. Anybody and everybody can use it for any kind of linguistic or non-linguistic purpose. For a specialised corpus, however, the question of target user is important. Since, each investigator or researcher has specific requirement, a corpus has to be designed accordingly. For instance, a person who is working on developing tools for machine translation will require a parallel corpus rather than a general corpus. Similarly, a person who is working on the comparative studies between two or more languages will require a comparable corpus rather than a monitor corpus. In the following list (Table 1) I have summed up the type of corpus users and their needs with regard to the type of corpus.

Target users	**Corpus**
Descriptive linguists	General, written, and speech corpus
NLP and LT people	General, monitor, parallel, spoken, aligned corpus
Speech technology people	Speech corpus and spoken corpus
Lexicographers and terminologists	General, monitor, specialised, reference, opportunistic corpus
Dialogue researchers	Speech, spoken, annotated, specialised corpus
Sociolinguistics	General, written, speech, monitor corpus
Psycholinguistics	Specialised, speech, written corpus
Historians	Literary, diachronic corpus
Social scientists	General, speech, written and special corpus
Comparative linguists	Bilingual, multilingual, parallel, comparable corpus
MT specialists	Bilingual, multilingual, parallel, comparable, annotated corpus
Information retrieval specialists	General, monitor, and annotated corpus
Tagging, processing and parsing specialists	Annotated, monitor, written, spoken, general corpus
Core-grammar designer	Comparable, bilingual, and general corpus
Word-Sense disambiguation worker	Annotated, monitor, written, spoken, general corpus
Teachers and students	Learner, monitor, and general corpus
Linguists	All types of corpus

Table 1: Type of corpus users and their needs with regard to the type of corpus

6.5 Selection of Time-span

Language changes with time. So determination of particular time-span is required to capture the features of a language or a variety reflected within a specific time span. A corpus should attempt to capture a particular period of time with a clear time indicator. For instance the text materials produced and published between the year 1981 and 1995 are included in the TDIL corpus with an assumption that the data will sufficiently represent the nature and character of use of the language as noted within this time-span. It is also hoped that this database will provide faithful information about the changes taking place within this period.

6.6 Selection of Texts Type

An important issue in text corpus generation is to determine if it will contain written texts of all types. Most of the corpora incline towards written texts of standard writings. The aim of a general corpus is to identify what are the central (common) and typical (special) to a language. Therefore, it is sensible to build a corpus with all kinds of texts of contemporary writings where a measured and proportional representation will suffice. To be realistic we should include works of the mass of ordinary writers along with works of established and well-known writers. Thus, a corpus will contain a collection of texts taken from all possible branches of human knowledge. Here the writings of highly reputed authors as well as of little-known writers are of equal importance.

The catalogues and lists of publications of different publishing houses need to be consulted for collection of source materials (e.g., books, newspapers, magazines, journals, etc.) for text selection and collection. A corpus which gathers texts from various sources and disciplines where individuality of particular source is made obscured is broadly heterogeneous in nature. In essence, diversity is the safeguard to corpus against any kind of skewed text representation.

The TDIL Bengali corpus contains texts from Literature (20%), Fine Arts (5%), Social Sciences (15%), Natural Sciences (15%), Commerce (10%), Mass media (30%), and Translation (05%). Each category has some sub-categories. E.g., *Literature* includes novels, short stories, essays etc.; *Fine Arts* includes paintings, drawings, music, sculpture etc.; *Social Science* includes philosophy, history, education etc.; *Natural Science* includes physics, chemistry, mathematics, geography etc.; *Mass Media* includes newspapers, magazines, posters, notices, advertisements etc.; *Commerce* includes accountancy, banking etc., and *Translation* includes all the subjects translated into Bengali.

6.7 Method of Data Sampling

Text data has to be sampled and selected from the collected source materials according to the need on an investigator. Sampling of data can be random, regular, or selective. There are various ways for data sampling to ensure maximum representativeness of texts in a corpus. We must clearly define the kind of language we want to study before we define the sampling procedures for the corpus. However, random sampling is a good technique that can save a corpus from being skewed and unrepresentative. We have used this technique for developing the TDIL corpus of Indian languages. This is a standard technique which is widely used in many areas of natural and social sciences.

Another approach is to define a sampling frame based on the nature of study. Designers of the *Brown Corpus*, for example, adopted this. They used all the books, periodicals, magazines, journals and other texts published in a particular year. This written corpus was made up with texts of different genres such as newspaper reports, romantic fictions, legal statutes, scientific writings, social sciences, technical reports, and so on.

Another reliable way is to use the complete bibliographical index available in a language. For instance, the *British National Bibliography,* and the *Willing's Press Guide* are used for generation of the *Lancaster-Oslo-Bergen* corpus.

6.8 Methods of Data Input

- **Data from electronic sources:** In this process texts from newspapers, journals, magazines, books etc. are included if these are found in electronic form.

- **Data from the websites:** This includes texts from web pages, web sites, and home pages.

- **Data from e-mails:** Electronic typewriting, e-mails, etc. are also used as source of data.

- **Machine reading of text:** It converts printed texts into machine-readable form by way of optical character recognition (OCR) system. Using this method, printed materials are quickly entered into a corpus.

- **Manual data input:** It is done through typing texts in computer. This is the best means for data collection from hand-written materials, transcriptions of spoken texts, and old manuscripts.

The process of data input is indirectly based on the method of text sampling. We can use two pages after every ten pages from a book. This makes a corpus best representative of data stored in a physical text. For instance, if a book has several chapters, each chapter containing different subject matters written by different writers, then the text samples collected in this process from all chapters are properly represented.

Each text file should have a Header which contains metadata – the physical information about the texts such as genre of the text (e.g., literature, science, commerce, technology, engineering, etc.), type of text (e.g., literature, story, travelogue, humour, etc), sub-type of text (e.g., fiction, historical, social, biographical, science fiction, etc.), name of book, name of the author(s), name of the editor(s), year of publication, edition number, name of the publisher, place of publication, number of pages taken for input, etc. This information is required for maintaining records and dissolving copyright problems.

It is also advantageous to keep detailed records of the materials so that the texts are identified on grounds other than those, which are selected as formatives of corpus. Information whether the text is a piece of fiction or non-fiction, book, journal or newspaper, formal or informal etc. are useful for both linguistic and non-linguistic studies.

At time of input of text, the original text of the physical source must be kept unchanged. After a paragraph is entered, one blank line should be given before a new paragraph starts. When texts are collected in a random sampling manner, a unique mark or flag is needed to be posted at the beginning of a new sample of text.

6.9 Hardware Requirement

For developing the TDIL corpus we used a Personal Computer (PC) with a GIST or Transcript Card (TC), a software, namely the Script Processor (SP), a monitor, one conventional computer key-board, a multilingual printer, and some floppy diskettes. Text files are developed with the help of the TC installed in the PC.

This allows the display of various Indian scripts on computer screen in a very convenient manner. The codes for various characters used for the Indian scripts are standardised by the *Bureau of Indian Standards, Govt. of India.* With installation of this card inside a PC, one can access almost the entire range of text-oriented application packages. One can also input and retrieve data in Indian languages with the help of this card. The software also provides a choice between two operational display modes on the monitor: one in conventional English mode, and the other in Indian multilingual mode.

6.10 Management of Data Files

Corpus data management is indeed a tedious task. It involves various related tasks such as holding, storing, processing, screening, and retrieving information from the corpus, which require utmost care and sincerity on the part of the corpus creators. Once a corpus file is created and stored in computer,

one needs schemes for its regular maintenance and augmentation. There are always some errors to be corrected, some modifications to be made, and some adjustment to be made, and some improvements to be implemented.

The adaptation to new hardware and software technology and the change in requirements of the users are also to be taken care of. In addition to this, there has to be constant vigilance on the retrieval tasks as well as on the processing and analytic tools to be applied on the corpus.

At present, the computer technology is developed to such an extent that executing these tasks with full satisfaction is no more a dream. But this solicits for more care in handling of the digital databases, since we know that the more powerful weapon we have in our hands, the more care is needed in their use and application to avoid unwanted damages on the resources.

6.11 Corpus Sanitation

The process of corpus sanitation begins (i.e., text editing) after the texts are produced in electronic form. Generally, five types of error may occur at the time of manual data entry (Dash 2008: 145-160):

(a) Omission of character,
(b) Addition of character,
(c) Repetition of character,
(d) Substitution of character, and
(e) Transposition of character.

To remove spelling errors, we need to thoroughly check the corpus and compare it with the actual physical data source, and do manual corrections. Care has to be taken to ensure that the spelling of words used in the corpus must resemble with the spelling of words used in the source texts. Also, it has to be checked if words are changed, repeated or omitted, punctuation marks are properly used, lines are properly maintained, and separate paragraphs are made for each text.

Besides error correction, we have to verify the omission of foreign words, quotations, dialectal forms, etc. after generation of a corpus. The naturalised foreign words are, however, allowed to enter into the corpus. Others should be omitted. Dialectal variations are allowed. Punctuation marks and transliterated words are faithfully reproduced. Usually, books on natural and social sciences contain more foreign words, phrases and sentences than books of stories or fiction.

Similarly, quotations from other languages, poems, songs, mathematical expressions, chemical formulae, geometric diagrams, images, tables, pictures, figures, flow-charts and similar symbolic representations of the source texts are not entered into corpus. All kinds of processing and reference works become easier and authentic if corpus is properly edited and errors are removed.

6.12 Problems of Copyright

To be in the safe side we need clearance of copyright from all the copyright holders (i.e., publishers and/or authors, and speakers for spoken texts). Copyright laws are quite complicated. There is very little which is obviously right or wrong, and legal or illegal. Moreover, copyright problems differ in various countries. If one uses the material only for personal use, then there is no problem. This is fine not only for a single individual but also for a group of scholars who are working together on some areas of research and investigation.

So long it is not directly used for commercial purpose there is no problem to get copy-right permission for the texts to be stored in a corpus. However, using materials one can generate new tools and systems to commercialise. In that case also the copyright is not violated. The reformed generation of output provides safeguards against possible attacks from the copyright holders. But in case of direct commercial work, we must have prior permission from the legal copyright holders.

7. Corpus Processing

The need for corpus processing arises after the generation of a corpus. We need to devise systems, tools, techniques, and software for accessing the language data and for extracting relevant information from the corpus. Corpus processing is indispensable not only for mainstream linguistic research and development activities but also for language technology works.

There are various corpus processing techniques, such as statistical analysis, concordance, lexical collocation, key-word search, local-word-grouping, lemmatisation, morphological processing and generation, chunking, word processing, parts-of-speech tagging, annotation, parsing, etc. It has been observed that the results obtained from corpus processing often contradict intuitions about a language and its properties.

There are many corpus processing software available for English, French, German, and similar other languages. For the Indian languages, however, there are only a few. We need to design corpus processing tools for our own languages keeping the nature of Indian languages in mind.

Here I discuss in brief some well-known corpus processing techniques and tools actively used in English and other European languages. I describe these with close reference to English. Reference to Bengali and other Indian languages are made as and when necessary.

7.1 Frequency Count

Linguistics is a subject, which has a long relationship with statistics and mathematics. Mathematical linguistics, computational linguistics, corpus linguistics, applied linguistics, forensic linguistics, stylometrics, etc. requires different statistical and quantitative results obtained from natural language corpus. Without adequate knowledge of statistical information about different properties of a language we can make mistakes in handling linguistic data as well as in observation (Yule 1964: 10)

Corpus can be subject to both *quantitative* and *qualitative* analysis. In quantitative analysis we classify different linguistic properties or features of a particular language, count them and even construct more complex statistical models in an attempt to explain what is observed. It allows us to discover which phenomena are likely to be the genuine reflections of the behaviours of a language or a variety, and which are merely chance occurrences.

On the other hand, qualitative analysis aims at providing a complete and detailed description of the observed phenomena obtained from corpus by quantitative analysis. It allows finer distinctions to be drawn because it is not necessary to shoehorn data into a finite set of classifications.

Both quantitative and qualitative analyses have something to contribute to corpus study. However, results obtained from quantitative analysis are less rich than results obtained from qualitative analysis. There are different statistical approaches to achieve this goal.

Simple *descriptive statistical approach* enables us to summarise the most important properties of observed data. *Inferential statistical approach* uses information from *descriptive statistical approach* to answer the questions or to formulate hypothesis. *Evaluative statistical approach* enables to test whether a hypothesis is supported by evidence in data, and how the mathematical model or theoretical distribution of data relates to reality (Oakes 1998: 1). To perform comparisons one can also apply the *multivariate statistical techniques* (e.g., *Factor Analysis, Multidimensional Scaling, Cluster Analysis, Log-linear Models* etc.) to extract hidden patterns from raw frequency data obtained from a corpus.

7.2 Word Sorting

Words stored in a corpus can be sorted in two ways. The process of *numerical sorting* is the most straightforward approach to work with quantitative data. Here items are classified according to a particular scheme, and an arithmetical count is made on the number of items within the texts, which belong to each class in the scheme. Information available from simple frequency counts are rendered either in alphabetical or in numerical order. Both the lists can again be arranged in ascending or in descending order according to our requirement. Any one who is studying a text will like to know how

often each different item occurs in it. A frequency list of words is a set of clues to texts. By examining the list we get an idea about the structure of text and can plan an investigation accordingly.

On the other hand, the *alphabetically sorted word list* is used for simple general references. A frequency list in alphabetical order plays a secondary role because it is used only when there is a need to check frequency of a particular item. However, it is useful as an object of study as it is often helpful in formulating hypotheses to be tested, and checking assumptions that have been made before hand Kjellmer (1984).

Before we initiate frequency counting on the Indian language corpora, we need to take decisions about the process of dealing with the characters, words, idioms, phrases, clauses and sentences used in the corpus. These will restrain us from false observations and wrong deductions about the various linguistic properties of the languages.

7.3 Concordance

The process of concordance refers to making an index to words used in a corpus. It is a collection of occurrences of words, each in its own textual environment. Each word is indexed with reference to the place of its occurrences in the texts. It is indispensable strategy because it gives access to many important language patterns in the texts. It provides information not accessible via intuitions.

There are some concordance software available for analysing corpus, e.g., *MonoConc* for sorting and frequency, *ParaConc* for parallel texts processing, *Conc* for sorting and frequency counting, *Free Text* for processing, sorting, etc. Concordance is most frequently used for lexicographical works and for language teaching. One can use it to search out single as well as multiword strings, words, phrases, idioms, proverbs, etc. It is also used to study lexical, semantic, syntactic patterns, text patterns, genre studies, and style patterns of texts (Barlow 1996). It is an excellent tool for investigating words and morphemes, which are polysemous and have multiple functions in a language.

7.4 Lexical Collocation

Method of collocation on words helps to understand the role and position of words in a text. Our traditional linguistic descriptions and hypotheses are challenged by new evidences accumulated from corpus through collocation.

Recent work on lexicography (Ooi 1998) shows that for many common words, the most frequent meaning is not the one that first comes to our mind and that takes place in dictionaries. Thus, it helps to determine which pairs of words have substantial collocational relationship between themselves. It compares the probabilities of two words occurring together as an event with the probability that they are simply the result of chance. For each pair of words, a score is given – the higher is the score the greater is the collocational ability of the words. It enables to extract multiword units from a corpus to use in lexicography and technical translation. It also helps to group similar words together to identify sense variations (e.g., *river bank* = landscape, but *investment in bank* = financial use.)

It helps in discriminating differences in usage between the words, which are similar in meaning. For instance, *strong* collocates with *motherly, showings, believer, currents, supporter, odour,* etc. while *powerful* collocates with *tool, minority, neighbour, symbol, figure, weapon, post* etc. (Biber at al. 1998: 165). Such information about delicate differences in collocation between the two similar words has an important role in helping students to learn a language in a better way.

The information about collocations of various linguistic items (e.g., words, morphs, idioms, etc.) is important for dictionary writing, NLP works, and machine translation besides language education. However, it is not easy to determine which co-occurrence is a significant collocation, especially if one is not a native speaker of a language or a language variety.

7.5 Key-Word-In-Context (KWIC)

KWIC is widely used in corpus processing to look up each occurrence of particular words (similar to concordance). The word under investigation appears at the centre of each line, with extra space on both sides. The length of the context may be different for different purposes. It shows an environment

of two, three or four words on either side of the word at the centre. This pattern may vary according to one's need. At the time of analysis of words, phrases, and clauses it is agreed that additional context is needed for better understanding.

It is better to think KWIC as a text in itself, and examine frequency of words in the environment of the central word. It is not that all information is needed every time but we utilise information when we require. After analysing a corpus by KWIC we can formulate various objectives in linguistic description and devise procedures for pursuing these objectives.

KWIC helps to understand the importance of context, role of associative words, actual behaviour of words in contexts, actual environment of occurrence, and if any contextual restriction is present in the use of a word. For instance, KWIC on the *Bank of English* shows that the most frequently used verb in reflexive form is *find* followed by *see, show, present, manifest* and *consider* – all of which involve 'viewing' of a representation or proposition.

7.6 Local Word Grouping (LWG)

LWG is another type of text analysis or corpus processing, which throws light on the patterns of use of a group of words in the texts. It is useful where word order is important for determining semantic load of sentences, and where individual semantic load of a constituent is affected for the presence of another constituent. It provides information for dealing with the functional behaviour of constituents at the time of parsing, both in phrase and sentence level.

LWG on the BNC shows that verb *manifest* is mostly associated with the third person neuter reflexives, whereas *enjoy* occurs with all reflexive forms except neuter (Barlow 1996). It determines the distribution of verb forms (e.g., *amuse oneself, please oneself, lend oneself, remind oneself*, etc.) which are not very common in use but which have a special affinity for reflexive forms. Knowledge of such patterns is important in moving language learners from intermediate to more advanced levels of proficiency.

Using LWG we find that most of the non-finite verbs are followed by finite verbs, while nouns are mostly preceded by one or two adjectives and followed by suffixes and postpositions in Bengali. It helps to analyse the so called *verb groups* and *noun groups* from their local information in case of a language like Hindi. It provides clues for understanding their roles in phrases, clause, and sentences. Information from LWG helps to dissolve lexical ambiguity, which arises from the local association of various lexical items.

My experience with the Bengali corpus shows that finer shades of meaning are mostly conveyed by internal relation between the constituents along with their distributions in contexts (Dash 2003). For many compound nouns and verbs, meaning denoted by a particular association of words cannot be obtained from meanings of individual words.

7.7 Morphological Processing

Morphological processing involves automatic analysis of words used in corpus. It includes processing of single word units, double word units, multiword units, etc. The main objective is to identify a word in a piece of text, isolate it from its contextual environment, analyse its morphophoemic structure, obtain its original meaning, and define its syntactic role it plays in a text. Information obtained from word processing is valuable for word sense disambiguation, dictionary compilation, parsing, language learning, etc.

There are several word processors for English and other languages (Greene and Rubin 1971, Karttunen and Wittenburg 1983, Koskenniemi 1983, deHaan 1984, Garside 1987, Church et al. 1991, deRose 1991, Merialdo 1994). For processing Bengali words I used *Directed Acyclic Graph Based Approach* (Sengupta and Chaudhuri 1993), *Trie-structure Based Approach* (Chaudhuri, Dash, and Kundu 1997), and *GS_Morph Approach* (Sengupta 1999). The *Trie-structure Based Approach* is orthography-based to follow conventional spelling system. It uses no phonological rules. Linguistic knowledge behind this is simple, and implementation algorithm is straightforward that requires little programming skill. An example from Bengali:

Surface form	: baleichilaam
Word-class	: Finite Verb
Root part	: bal-
Suffix part	: -eichilaam
Aspect marker	: -e-
Particle marker	: -i- (emphatic)
Auxiliary marker	: -ch-
Tense marker	: -il(a) (past)
Person marker	: -aam (1st)
Honorific marker	: Null
Number marker	: Null (Sng./Pl.)
Meaning	: "I/we had said"

People working on native language can have better results since intuitive knowledge helps in finding out right root or suffix part from the inflected words, which may be beyond the grasp of non-native users.

7.8 Part-of-Speech Tagging

Certain types of linguistic annotation, which involve attachment of special codes to words in order to indicate particular features, are often known as *tagging* rather than *annotation*. Codes, which are assigned to the features, are known as *tags***.**

The parts-of-speech tagging scheme tags a word with the part-of-speech it is used in a sentence. It is done at three stages: (a) pre-editing, (b) automatic tag assignment, and (c) manual post-editing.

In pre-editing stage, corpus is converted to a suitable format to assign a part-of-speech tag to each word or word combination. Because of orthographic similarity, one word may have several possible POS tags. After initial assignment of possible POS, words are manually corrected to disambiguate words in texts.

For Bengali I suggest the following tagset for the words belonging to major lexical classes: Noun [NN], Pronoun [PN], Adjective [ADJ], Finite Verb [FV], Adverb [ADV], Non-finite verb [NFV], Postposition [PP], Indeclinable [IND], and Reduplication [RDP].

The second level of POS tagging is called *grammatical tagging*. It is a more comprehensive annotation system, which assigns all grammatical information to each word in corpus. The scheme is used to distinguish the inflected words necessary for skeleton parsing. To solve the problems related to morphological, lexical, semantic, and syntactic analysis we urgently need a corpus, which includes grammatical information for each word occurring in the texts.

7.9 Word Sense Tagging

Word sense tagging or semantic tagging accepts words as input text, which is tagged with POS and grammatical tagging. After automatic tag assignments are over, manual post-editing is done to ensure that each word carries correct semantic classification. Following example of word sense tagging is taken from a project at *Lancaster University*, UK.

PPIS1	I	Z8
VV0	like	E2+
AT1	a	Z5
JJ	particular	A4.2+
NN1	shade	O4.3
IO	of	Z5
NN1	lipstick	B4

In this table the text is read downwards, with grammatical tags on the left, and the word sense tags on the right. Semantic tags are composed of an upper case letter indicating general discourse field, a

digit indicating a first subdivision of the field, a decimal point followed by a further digit to indicate a finer subdivision, one or more 'pluses' or 'minuses' to indicate a positive or negative position on a semantic scale, etc.

For example, A4.2+ indicates a word in category 'general and abstract words' (A), subcategory 'classification' (A4), sub-subcategory 'particular and general' (A4.2), and 'particular' as opposed to 'general' (A4.2+). Likewise, E2+ belongs to category 'emotional states, actions, events and processes' (E), subcategory 'liking and disliking' (E2), and refers to 'liking' rather than 'disliking' (E2+).

7.10 Results of Tagging

After completing tagging on a small Bengali corpus, I have found three types of word: rightly tagged words, ambiguously tagged words, and untagged words.

(a) **Rightly tagged words:** Most of the PNs, NFVs, FVs, INDs, inflected NNs, ADVs, and ADJs are rightly tagged. Indeclinables are tagged with string matching. Most of the PNs and NNs are rightly tagged since root lexicon contains roots and suffix lexicon contains suffixes. Moreover, algorithm between root and suffix are agreed. Most FVs and NFVs are tagged rightly because root lexicon contains roots and suffix lexicon contains suffixes. Moreover, algorithm between root and suffix are agreed. Some ADJs and ADVs followed in the same way.

(b) **Untagged words:** Some words in the corpus are not tagged at all. Some FVs and NFVs are not tagged because root lexicon does not have the root forms. For NNs, PNs, and ADJs, either the suffix list is not exhaustive or the root part has failed to match with the root form stored in root lexicon. For indeclinable, the insufficiency in the lexicon dictionary. For ADVs, failure is caused for the space in between the words. Moreover, insufficiency in the list of adverbial root and suffix lexicon is also a reason. Some NNs are used as FVs in texts. Proper names, transliterated foreign words, dialectal forms are mostly untagged due to their absence in root lexicon. These are mostly technical problems, which can be removed by augmentation of respective root and suffix lexicon and modification of the matching algorithms.

(c) **Ambiguously tagged words:** Ambiguity at lexical level is quite common in all natural languages. A single lexical item can convey multiple items, sense, events, and ideas depending on the context of its uses. Efficiency and adequacy of a word processor comes from the way it handles lexical ambiguities. At POS tagging, ambiguity generally takes place at lexical level. *Lexical ambiguity* is caused because most of lexical items (FVs, NNs, ADJs, etc.) allow more than one reading or sense variation. These readings differ in terms of sub-categorisation features, selectional features, syntactic property, semantic property, idiomatic reading, figurative use, and so on (Sinclair 1991: 104-105).

After POS tagging, the Bengali corpus gives two types of ambiguity. *Structural ambiguity* is caused mostly for non-inflected words (e.g., *mat* 'opinion", *mat* 'like"), *caal* "rice" and *caal* "move", etc.), where root is homographic in form to belong to different lexical categories. Structural ambiguity is also noted in some inflected words (e.g., *kare* "havig done", *kare* "in hand", *bale* "having said" and *bale* "on a ball", *karaate* "to make other to do" and *karaate* "in saw", etc.) due to similarity both in root and suffix part. At the level of context-dependent parsing, such ambiguities can be dissolved.

Sequential ambiguity is mostly caused due to the presence of immediately following word (W_2), which if processed with the preceding word (W_1), produces a meaning which is different from their respective independent meanings. For instance, if *bishes* and *bhaabe* are processed in isolation, *bishes* means "specific" while *bhaabe* means "s(he) thinks". However, if these are processed together they will mean "specially", which is grossly different from their respective individual and independent meaning. To resolve sequential ambiguity, it is better to apply the method of delayed processing based on the KWIC or the LWG.

7.11 Lemmatisation

The process of lemmatisation is related to identification of inflected words used in a piece of text, and reducing to their respective lexemes or lemma form. It allows researchers to extract and examine all the variants of particular LEMMA without having to input all the possible variants, and to produce frequency and distribution information for the lemma.

It is useful for language teaching where the learners are trained to identify the total number of surface forms of a lemma. It is used to know which are inflected, how many times these are inflected, and in which way these are inflected, and so on. A part of the *Brown Corpus* contains the lemmatised forms of words along with all lexical and grammatical information. No Indian language corpus has been put to lemmatisation, as yet.

7.12 Annotation

Apart from pure texts, corpus is provided with additional linguistic information, known as annotation. Information is of different nature, such as part-of-speech, prosodic, semantic, anaphoric, discoursal annotation, etc. Annotated corpus is a very useful tool for research. Grammatically tagged corpus is the most common form of annotated corpus where words are assigned a word class. The *Brown Corpus*, *LOB Corpus*, and *BNC* are grammatically annotated, the *LLC* is prosodically annotated, while the *Susanne Corpus* is syntactically annotated. We are yet to start the work of annotation on the Indian language corpora.

(a) **Part of speech annotation:** In part-of-speech annotation the aim is to assign to each lexical unit in the text a code indicating its part-of-speech. It increases specificity of data retrieval from a corpus, and helps in syntactic parsing, and semantic field annotation. It allows us to distinguish between the homographs.

(b) **Anaphoric annotation:** In anaphoric annotation scheme, all pronouns and noun phrases are co-indexed within a broad framework of cohesion. Here different types of anaphora are listed and sorted. Such an annotation scheme is used for studying and testing mechanisms like pronoun resolution. It is important for text understanding and machine translation.

(c) **Prosodic Annotation:** In prosodic annotation the patterns of intonation, stress and pauses in speech are indicated. It is a more difficult type of annotation because prosody is considerably more impressionistic in nature than other linguistic levels. It requires careful listening by a trained ear. The *Lancaster/IBM Spoken English Corpus* is a prosodically annotated corpus where stressed syllables with and without independent pitch movement are marked with different symbols. All unstressed syllables, whose pitch is predictable from tone marks of surrounding accented syllables, are left unmarked.

(d) **Semantic Annotation:** In semantic annotation either the semantic features of words in a text (essentially, annotation of word senses) or the semantic relationships between the words in text (e.g., agents or patients of particular actions) are marked. There is no unanimously agreed norm about which semantic features ought to be annotated. Some propose (Garside, Leech, and McEnery 1997) to use *Roget's Thesaurus* where words are organised into general semantic categories. Such annotation scheme is designed to apply to both open-class (content words) and closed class of words, as well as proper nouns, which are marked by a tag and set aside from statistical analysis.

(e) **Discoursal Annotation:** In discoursal annotation a corpus is annotated at the levels of text and discourse, and is used in linguistic analysis. Despite its potential role in analysis of discourse this kind of annotation has never been widely used, possibly because linguistic categories are context-dependent, and their identification in texts is a greater source of dispute than other forms of

linguistic phenomena. Some have (Stenström and Andersen 1996) annotated the *London-Lund Spoken Corpus* with 16 discourse tags to observe trends in teenage talk.

7.13 Parsing

Parsing is actually related to the automatic analysis of texts according to a grammar (Barnbrook 1998: 170). Technically, it is used to refer to practice of assigning syntactic structure to a text (McEnery and Wilson 1996: 178). It is usually performed after the basic morphosyntactic categories have been identified within a text. Based on different grammars (e.g., dependency grammar, context free phrase structure grammar, systematic functional grammar, extended affix grammar, etc.) parsing brings these morphosyntactic categories into higher level syntactic relationships with one another. Sentence level parsing involves automatic context-based as well as context-free syntactic analysis using information acquired from word-level processing.

A parsed corpus is known as *treebank*, because it alludes to tree diagrams used in parsing. The visual diagram of tree structure is rarely found in corpus annotation. Generally, identical information is represented using sets of labelled brackets. Thus, *Pearl sat on a chair* will appear in treebank in the following way:

[S[NP Pearl_NP1 NP][VP sat_VVD [PP on_II [NP a_AT1 chair_NN1 NP] PP] VP] S]

where morpho-syntactic information is attached to words by underscore characters while constituents are indicated by opening and closing square brackets annotated at the beginning and end with phrase type e.g. [S ... S].

Not all parsing systems are similar. The main differences are: (i) the number of constituent types, which a system employs, and (ii) the way in which constituent types are allowed to combine with each other. However, despite these differences, majority of parsing schemes are based on a form of context-free phrase structure grammar.

Within this system, a *full parsing* scheme aims at providing a detailed analysis of sentence structure, while a *skeleton parsing scheme* tends to use a less finely distinguished set of syntactic constituent types and ignores internal structure of certain constituent types. Parsing is most often post-edited by human analysts because automatic parsing has a lower success rate than the part-of-speech annotation.

The disadvantage of full manual parsing is inconsistency on behalf of analyst(s) engaged in parsing or editing corpus. To overcome this, more detailed guidelines are provided but even then ambiguities may occur when multiple interpretations are possible.

Treebanks are language resource that provides annotations of natural languages at various levels of structure: at word level, phrase level, sentence level, and sometimes at the level of function-argument structure. Treebanks have become crucially important for developing data-driven approaches to natural language processing, human language technologies, grammar extraction, and linguistic research in general. There are a few on-going projects on compilation of representative treebanks for many European and USA languages. Implementation of such system on Indian corpora requires more time and research.

Processing of corpus texts is of high importance to any language processing system that attempts to use natural language in some way or other. Advanced requirements of users raise need for efficient and widely applicable systems. Need for comprehensive processing capabilities has strong interface among theoretical, applied, and computational linguistics. Given the complexity of natural languages, it is always difficult for a machine to make accurate decisions about any property of a language. Therefore, occasional errors in processing should not be taken as a major road-block to research in language processing. An interactive computer program designed specifically for checking errors can make this process much faster and more reliable.

8. Utility of Corpus

Unless defined otherwise, let us consider that a corpus should possess all the properties mentioned in Section 3. In essence, a corpus is an empirical standard, which acts as a benchmark for validation of usage of linguistic properties found in a language. If one analyses a corpus database, one can retrieve the following information about a language or variety.

- Information about all the properties and components used in a language, e.g., sounds, phonemes, intonation, letters, punctuations, morphemes, words, stems, bases, lemmas, compounds, phrases, idioms, set phrases, reduplications, proverbs, clauses, sentences, etc.

- Grammatical and functional information of letters, graphemes, allograpohs, morphemes, words, phrases, sentences, idiomatic expressions, proverbs, etc. relating to their structure, composition, patterns of using affixes and inflections, patterns of constituent structure, contexts of use, usage patterns, variations of contexts, etc.

- Usage-based information of letters, characters, phonemes, morphemes, words, compounds, phrases, sentences, etc. relating their descriptive, stylistic, metaphorical, allegorical, idiomatic, and figurative usages, etc.

- Extralinguistic information relating to time, place, situation, and agent of language events, social-cultural backgrounds of linguistic acts, life and living of target speech community, discourse and pragmatics, as well as of the world knowledge of the language users at large.

It is understandable that developing a corpus in accordance with these pre-conditions mentioned above is really a tough task. However, we can simplify the task to some extent if we redefine the entire concept of corpus generation based on object-oriented and work-specific needs. Since it is known that all types of corpus should not follow the same set of designing and composition principles we can have liberty to design a corpus keeping in mind the works we are planning to do with it (Dash 208: 47). The underlying proposition is that the general principles and conditions of corpus generation may vary depending on the purpose of a corpus developer or a user.

Corpus linguistics is, however, not the same thing as obtaining language databases through the use of computer. It is the processing and analysis of the data stored within a corpus. The main task of a corpus linguist is not to gather databases, but to analyse these. Computer is a useful, and sometimes indispensable, tool for carrying out these activities.

9. Use of Corpus

There are a number of areas where language corpus is directly used as in *language description, study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, lexicography, discourse, pragmatics, language teaching, language planning, sociolinguistics, psycholinguistics, semiotics, cognitive linguistics, computational linguistics* — to mention a few. In fact, there is hardly any area of linguistics where corpus has not found its utility. This has been possible due to great possibilities offered by computer in collecting, storing, and processing natural language databases. The availability of computers and machine-readable corpora has made it possible to get data quickly and easily and also to have this data presented in a format suitable for analysis.

- **Corpus as knowledge resource:** corpus is used for developing multilingual libraries, designing course books for language teaching, compiling monolingual dictionaries (printed and electronic), developing bilingual dictionaries (printed and electronic), multilingual dictionaries (printed and electronic), monolingual thesaurus (printed and electronic version), various reference materials (printed and electronic version), developing machine readable dictionaries (MRDs), developing multilingual lexical resources, electronic dictionary (easily portable, can be duplicated as many

copies as needed, can be modified easily for newer versions, can be customised according to need of users, can be ready and accessed easily, more durable than printed dictionary, etc.).

- **Corpus in language technology:** corpus is used for designing tools and systems for word processing, spelling checking, text editing, morphological processing, sentence parsing, frequency counting, item-search, text summarisation, text annotation, information retrieval, concordance, word sense disambiguation, WordNet (synset), semantic web, Semantic Net, Parts-of-Speech Tagging, Local Word Grouping, etc.

- **Corpus for translation support systems:** corpus is used for language resource access systems, Machine translation systems, multilingual information access systems, and cross-language information retrieval systems, etc.

- **Corpus for human-machine interface systems:** corpus is used for OCR, voice recognition, text-to-speech, E-learning, on-line teaching, e-text preparation, question-answering, computer-assisted language education, computer-aided instruction, e-governance, etc.

- **Corpus in speech technology:** Speech corpus is used to develop general framework for speech technology, phonetic, lexical, and pronunciation variability in dialectal versions, automatic speech recognition, automatic speech synthesis, automatic speech processing, speaker identification, repairing speech disorders, and forensic linguistics, etc.

- **Corpus in mainstream linguistics:** corpus is used for language description, lexicography, lexicology, paribhasa formation, grammar writing, semantic study, language learning, dialect study, sociolinguistics, psycholinguistics, stylistics, bilingual dictionary, extraction, translation equivalents, generation of terminology databank, lexical selection restriction, dissolving lexical ambiguity, grammatical mapping, semiotics, pragmatic and discourse study, etc.

10. Potential Corpus Users

Let us visualise that language corpora can be used by (a) language specialists who are interested in the language of the texts, (b) content specialists who are interested in the content of the texts, and (c) media specialists who are interested in corpus as test bed for electronic devices.

(a) Among the **language specialists,** *Lexicographers* consults corpus for information on actual and specific use of words, lexemes, phrases, idioms, etc. They process corpus in order to develop lexical databases, dictionaries, thesauruses, and reference materials. The *terminologists* and *technical writers* use corpus database to standardise the technical terminology as well as increase terminology databases. *Theoreticians* use corpus as a large body of mass representation of facts of language. For them corpus yields data on relative frequency of phenomena of all kinds, and provide scope for verification of evidences of their own or their informants. *Applied linguists* use corpus in language teaching, since it supplies resources for extracting and studying language with authority of attested use.

(b) Among the **content specialists,** *historians* are able to track development of opinions and ideas through the study of words, phrases, and sentences that refer to them in the corpus. They also use dated and analysed corpus to discover implicit time- and place-stamps, which they use to identify documents, whose origin is obscured. *Literary critics* use corpus for research into stylometrics, since statistical analysis of word-use plays crucial role in determining ascription of dubious works to known authors. Such techniques become more effective when linguists discover significant features present at a higher level than individual words. Besides acting as a mass training ground for techniques, it is used as resource for statistical information on the differences of style characterising different groups, identified by age, sex, period, country of origin, etc. *Sociologists*

use corpus in similar fashion to characterise different groups belonging to different class, race, creed, ethnicity, etc.

(c) Among the **media specialists,** *information retrievers* use corpus to devise mechanisms for extracting appropriate information from bodies of text to build up linguistic knowledgebase, find information of items for indexing, and summarise important content of texts. *Computational linguists* use corpus to integrate their works with statistical regularities found in corpus, which work as an important key to analyse and process language. Also, corpus, as a source of data and knowledgebase is used for testing presence or absence of regularities in language, since statistical techniques become more effective when they work on outputs of some grammatically analysed corpora. *Machine translators* may access corpus to extract necessary and relevant linguistic information and verify efficiency of their systems, since corpus makes significant contribution to enhance actual capability of systems. Moreover, domain specific corpus enables systems to adopt self-organising approach to supplement traditional knowledge-based approaches. The *language processing people* benefit more and more from the development of corpus of various types, since both raw and annotated corpora are used in a large scale for developing language processors. It is suffice to say that corpus is a beneficial resource for all – including researchers, technologists, writers, lexicographers, academicians, teachers, students, language learners, scholars, publishers, and others.

11. Limitations of Corpus

(a) **Lack of linguistic generativity:** Chomsky and his supporters have strongly criticised the value of corpus in linguistic research. At the University of Texas in 1958, he argued, "any natural corpus will be skewed. Some sentences won't occur because they are obvious; others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list". Generativists argue that corpus cannot provide evidence for linguistic innateness. By virtue of its structure and content, it only can represent the linguistic 'performances' but does not reflect on the linguistic 'competence' and 'generitivity' of the users. A corpus, which records only the examples of performance, cannot be useful to linguists, who seek to understand the tacit, internalised knowledge of language rather than the external evidences of language use on various contexts.

(b) **Technical difficulties:** Corpus building is a large scale, multidirectional, enterprising work. It is a complex, time-consuming, error-prone, and expensive task. The whole enterprise requires an efficient data processing system, which may not available to all, particularly in a country like India. Linguists need to be trained in computer use and data handing. It is a troublesome task. Unlike linguists of other countries, Indian linguists are not eager to take up computer into their stride. Computer scientists, on the other hand, are also not enthusiastic to work with the linguists in tandem. The gap is wide apart. Let us hope for a mutual co-operational interface to develop between the two groups in near future.

(c) **Lack of texts from dialogues:** Present day corpus fails to consider the impromptu, non-prepared dialogues taking place spontaneously in daily linguistic exercises. Absence of texts from dialogic interactions makes a corpus cripple lacking in the aspect of spontaneity, a valuable trait of human language. Corpus, either in spoken or written form, is actually a database detached from the actual context of language use. Detachment from the contexts makes a corpus (corpse + carcass) a dead database, which is devoid of many properties of living dialogic interactions, discourse, and pragmatics. It fails to reveal the real purpose underlying a linguistic negotiation (a difficult action game), identify the language-in-use, determine the verbal actions involved within the dialogues, describe the background where from the interlocutors derive cognitive and perceptual means of communication.

(d) **Lack of information from visual elements:** Corpus does not contain graphs, tables, pictures, diagrams, figures, images, formulae and similar other visual elements, which are often used in a piece of text for proper cognition and understanding. A corpus devoid of such visual elements is bound to lose much of its information.

(e) **Other limitations:** Corpus creation and research works are unreasonably tilted towards written texts, which reduce importance of speech. In reality, however, speech represents our language in a more reliable fashion than writing. The complexities of speech corpus generation make it a rare commodity. Thus, easy availability of text corpus and the lack of speech corpus inspire people to turn towards the text corpus. However, this does not imply that speech corpus has lost is prime position in corpus linguistics research. Moreover, language stored in corpus fails to highlight the social, evocative, and historical aspects of language. Corpus cannot define why a particular dialect is used as the standard one, how dialectal differences play decisive roles to establish and maintain group identity, how idiolect determines one's power, position and status in society, how language differs depending on domains, registers, etc. Corpus also fails to ventilate how certain emotions are evoked by certain poetic texts, songs and literature; how world knowledge and context play important roles to determine intended meaning of an utterance; how language evolve, divide, and merge with the change of time and society, etc.

12. Suggested Reading

Aarts, J. and Meijs, W. (Eds.) 1984. *Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research*. Amsterdam-Atlanta, GA.: Rodopi.

Aarts, J. and Meijs, W. (Eds.) 1986. *Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora.* Amsterdam-Atlanta, GA.: Rodopi.

Aijmer, K. and Altenberg, B. (Eds.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Aston, G. (Ed.) 2004. *Learning with Corpora*. Cambridge: Cambridge University Press.

Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.

Baker, M., Gill, F. and Tognini-Bonelli, E. (Eds.) 1993. *Text and Technology: In honour of John Sinclair*. Philadelphia: John Benjamins.

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.

Bernadini, S. 2000. *Competence, Capacity, Corpora*. Bologna: CLUEB.

BIBER, D., CONRAD, S. and REPPEN, R. 1998. *Corpus linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Boguraev, B. and Pustejvsky, J. (Eds.) 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, Mass.: MIT Press.

Botley, S.P., A.M. McEnery, and A. Wilson (Eds.) 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA: Rodopi.

Bouillon, P. and Busa, F. (Eds.) 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press.

Butler, C.S. (Ed.) 1992. *Computers and Written Texts*. Oxford: Blackwell Publishers.

Carter, R. and McCarthy, M. (Eds.) 1988. *Vocabulary and Language Teaching*. London: Longman.

Coleman, J. and Kay, C.J. (Eds.) 2000. *Lexicology, Semantics and Lexicography: Selected Papers from the 4th G.L. Brook Symposium.* Amsterdam/ Philadelphia: John Benjamins.

Collins, P. and Blair, D. (Eds.) 1989. *Australian English*. St. Lucia: University of Queensland Press.

Cuyckens, H. and Zawada, B. (Eds.) 2001. *Polysemy in Cognitive Linguistics.* Amsterdam/ Philadelphia: John Benjamins.

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

Dash, N.S. (2007) *Language Corpora and Applied Linguistics*. Kolkata: Sahitya Samsad.

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

Dash, N.S. (2009) *Corpus Linguistics: Past, Present and Future*. New Delhi: Mittal Publications.

Dash, N.S. (2009) *Corpus-based Analysis of the Bengali Language*. Saarbrucken, Germany: Verlag Dr Muller Publications.

Edwards, J.A. and Lampert, M.D. (Eds.) 1993. *Talking Data: Transcription and Coding in Discourse Research.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Francis, W.N. and Kucera, H. 1964. *Manual of information to accompany A standard Corpus of present-day edited American English.* Dept. of Linguistics, Brown University, USA.

Francis, W.N. and Kucera, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Fries, U., Mü, V. and Schneider, P. (Eds.) 1997. *From Aelfric to the New York Times*. Amsterdam: Rodopi.

Garside, R., Leech, G. and McEnery, A. (Eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Garside, R., Leech, G. and Sampson, G. (Eds.) 1987. *The Computational Analysis of English: A Corpus Based Approach.* London: Longman.

Gerbig, A. 1997. *Lexical and Grammatical Variation in a Corpus: A Computer-Assisted Study of Discourse on the Environment.* London: Peter Lang Publishing.

Ghadessy, M., Henry, A. and Roseberry, R.L. Eds. 2001. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam/ Philadelphia: John Benjamins.

Granger, S. and Tyson, S.P. (Eds.) 2003. *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Amsterdam: Rodopi.

Granger, S., Hung, J. and Tyson, S.P. Eds. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Greenbaum, S. (Ed.) 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon.

Greene, B. and Rubin, G. 1971. *Automatic Grammatical Tagging of English.* Technical Report. Department of Linguistics. Brown University, RI, USA.

Halliday, M.A.K. 1987. *Spoken and Written Modes of Meaning, Comprehending Oral and Written Language.* San Diego, CA: Academic Press.

Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Halteren, H.V. (Ed.) 1999. *Syntactic Word Class Tagging*. Dordrecht: Kluwer Academic Press.

Hickey, R. and Stanislaw, P. (Eds.) 1996. *Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak.* Vol. 2. Berlin: Mouton de Gruyter.

Hofland, K. and Johansson, S. 1982. *Word Frequencies in British and American* English. Bergen: Norway Computing Centre for the Humanities.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hutchins, W.J. 1986. *Machine Translation: Past, Present, and Future.* Chichester: Ellis Harwood.

Ilson, R.F. (Ed.) 1986. *Lexicography: An Emerging International Profession.* Manchester: Manchester University Press.

Jensen, J.T. 1990. *Morphology: Word Structure in Generative Grammar*. Amsterdam: John Benjamins.

Jespersen, O. 1909-1949. *Modern English Grammar on Historical Principles*. 7 Vols. London: Allen and Unwin.

Johansson, S. and Hofland, K. (Eds.) 1982. *Computer Corpora in English Language Research.* Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. and Stenström, A-B. (Eds.) 1991. *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.

Katamba, F. 1993. *Morphology*. London: Macmillan Press.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* New York: Addison-Wesley Longman Inc.

Kenny, A.J.P. 1982. *The Computation of Style*. Oxford: Pergamon Press.

Kettemann, C.B. and Marko, G. (Eds.) 2002. *Teaching and Learning by Doing Corpus Analysis. Language and Computers:* Studies in Practical Linguistics 42. Amsterdam-Atlanta, GA.: Rodopi.

Kilgariff, A. and J. Palmer (Eds.) 2000. *Computer and the Humanities: Special Issue on Word Sense Disambiguation.* Vol. 34. No.1. 2000.

Kirk, J.M. (Ed.) 2000. *Corpora Galore: Analyses and Techniques in Describing English.* Amsterdam; Atlanta, GA: Rodopi.

Kucera, H. and Francis, W.N. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Kytö, M., Ihalainen, O. and Rissanen, M. (Eds.) 1988. *Corpus Linguistics, hard and soft: Proceedings of the 8th International Conference on English Language Research on Computerised Corpora.* Amsterdam: Rodopi.

Lancashire, I., E. Carol, and C.F. Meyer (Eds.) 1997. *Synchronic Corpus Linguistics*. Bergen, Norway: ICAME.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Leech, G., G. Myers, and J. Thomas (Eds.) 1995. *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

Levy, M. 1997. *Computer Assisted Language Learning.* Oxford: Oxford University Press.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17th International Conference on English-Language Research Based on Computerized Corpora. Amsterdam: Rodopi.

Macwhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, N.J.: Lawrence Erlbaum.

Mair, C. and Hundt, M. (Eds.) 2000. *Corpus Linguistics and Linguistics Theory*. Amsterdam-Atlanta, GA: Rodopi.

McArthur, T. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

McCarthy, J. 1982. *Formal Problems in Semitic Phonology and Morphology*. New York: Garland.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics.* Cambridge: cambridge University press.

McEnery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mcenery, T., Rayson, P. and Wilson, A. (Eds.) 2002. *A Rainbow of Corpora: Corpus Linguistics and the Langauges of the World.* München: Lincom Europa.

Meyer, C.F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, G.A. 1951. *Language and Communication.* New York: McGraw-Hills.

Nelson, G., Wallis, S. and Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English.* Amsterdam/ Philadelphia: John Benjamins.

Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, V.B.Y. 1997. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press

Oostdijk, N. and deHann, P. (Eds.) 1994. *Corpus Based Research into Language*. Amsterdam-Atlanta, GA: Rodopi.

Partington, A. 1998. *Patterns and Meanings - Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins.

Percy, C., Meyer, C.F. and Lancashire, I. (Eds.) 1996. *Synchronic Corpus Linguistics.* Amsterdam-Atlanta, GA: Rodopi.

Peters, B.P., Collins, P. and Smith, A. (Eds.) 2002. *New Frontiers of Corpus Research*. Language and Computers. Amsterdam-Atlanta, GA: Rodopi.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Ravin, Y. and Leacock, C. (Eds.) 2000. *Ploysemy: Theoretical and Computational Approaches*. New York: Oxford University Press Inc.

Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Cambridge: Cambridge University Press.

Selting, M. and Couper-Kuhlen, E. (Eds.) 2001. *Studies in Interactional Linguistics*. Amsterdam/ Philadelphia: John Benjamins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Souter, C. and Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics.* Amsterdam: Rodopi.

Sperberg-Mcqueen, C.M. and Burnard, L. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford: ACH-ACL-ALLC Text Encoding Initiative.

Stenström, A-B, Andersen, G. and Hasund, I.K. 2002. *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings.* Amsterdam: John Benjamins.
Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell. Publishers.
Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.
Svartvik, J. (Ed.) 1990. *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund: Lund University Press.
Svartvik, J. (Ed.) 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 - Stockholm*, 4-8 August 1991. Berlin, New York: Mouton De Gruyter.
Tannen, D. (Ed.) 1982. *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, New Jersey: Ablex Publishing Corporation.
Thomas, J. and Short, M. (Eds.) 1996. *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London and New York: Addison Wesley Longman.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
Vera, D.E.J. (Ed.) 2002. *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics.* Amsterdam: Rodopi.
Véronis, J. (Ed.) 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.
Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (Eds.) 1997. *Teaching and Language Corpora.* London: Longman.
Young, S. and G. Bloothooft (Eds.) 1997. *Corpus-Based Methods in Language and Speech Processing.* Vol-II. Dordrecht: Kluwer Academic Press.

Dr Niladri Sekhar Dash
Assistant Professor
Linguistic Research Unit
Indian Statistical Institute
203, Barrackpore Trunk Road
Baranagar, Kolkata-700108
West Bengal, India
Email: niladri@isical.ac.in
Email: ns_dash@yahoo.com
Email: nisedash@gmail.com
http://www.isical.ac.in/~niladri