

WHY DO WE NEED TO DEVELOP CORPORA IN INDIAN LANGUAGES ?

NILADRI SEKHAR DASH AND BIDYUT BARAN CHAUDHURI
COMPUTER VISION AND PATTERN RECOGNITION UNIT
INDIAN STATISTICAL INSTITUTE
203, BARRACKPORE TRUNK ROAD
KOLKATA 700108, INDIA
Email: {niladri/bbc@isical.ac.in}

ABSTRACT

The easy availability of language corpora and their processing tools have opened up many new areas of language research which were unknown to us even a few decades ago. Language corpora, and the results obtained from them have put intuitive language study under strong challenge. In most cases, intuitive observation are proved to be wrong or inadequate while compared with the findings from corpora. Thus, corpora have proved their utility in empirical language analysis, theory making, as well as in theory modification which were missing in intuitive language study. However, this trend of corpus-based language research is yet to set its firm footing in India though there have been some sporadic attempts for developing corpora in Indian languages. We should realise that in a multilingual country like India we need to develop language corpora of various types not only to be at par with language related technology developed in other countries, but also to provide advanced resources and systems to our people for their education and research. With this information in background, this paper focuses on how various types of corpora can be designed in Indian languages, how they can be used for developing language processing tools, how developed systems and tools can be used in language-related activities, and how information acquired from corpora can be utilised in language research and education at various levels.

Key words: corpora, processing, grammars, machine translation, language technology, dictionary, sociolinguistics, language learning, lexicography, semantics, polysemy, stylistics, speech, etc.

1. INTRODUCTION

Language corpora usually represent a large collection of representative samples obtained from texts covering different varieties of language used in various domains of linguistic activities. Theoretically, corpora should be (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts. In fact, the term CORPUS can be derived from its features it implies:

C	: Compatible to computers
O	: Operational in research and application
R	: Representative of the source language
P	: Processable by both man and machine
U	: Unlimited in amount of data, and
S	: Synchronic in formation and representation

Unless defined otherwise, let us consider that corpora should possess all the properties mentioned above. Exception can be noted for historical corpora which are neither unlimited nor synchronic. However, such corpora are mostly confined within a specific peripheral zone of corpus linguistics having marginal importance in the whole gamut of empirical language research. By all means language corpora should:

- be good representative of a natural language
- be large and balanced to all disciplines
- try to capture almost all the linguistic features of a language
- be easily retrieved and used by the end-users
- be subjected to any kind of empirical investigation and analysis, and
- be good resource for developing sophisticated language processing tools

The issues related to corpora development and text processing may vary depending on the type of corpora and the purpose of their use. Issues related to speech corpora development may differ from issues related to written corpora development. While developing written corpora we have to take care of issues like determination of target users, choice of documents, selection of time-span, manner of data input, size of corpora, problem of copy-right, manner of sampling, problem of text screening, management of corpus file, editing of input data, processing of texts, analysis of text, etc. Similarly, while developing speech corpora we have to consider issues like selection of informants, propose of use, choice of settings, selection of speakers, manner of data collection, size of corpora, problem of transcription, type of data encoding, manner of data-sampling, management of data files, editing of input data, processing of texts, analysis of texts, etc. One can have good information on features of corpora, design and development, methods of text processing, types of text encoding, part-of-speech tagging, parsing, sense decoding, and other relevant issues of corpus linguistics from Aarts and Meijs (1984, 1986), Garside, Leech and Sampson (1987), Aijmer and Altenberg (1991), Johansson and Stenstrom (1991), Sinclair (1991), Svartvik (1992), Oostdijk and deHann (1994), McEnery and Wilson (1996), Percy, Meyer and Lancashire (1996), Thomas and Short (1996), Garside, Leech and McEnery (1997), Knowles, Williams and Taylor (1997), Lancashire, Carol and Meyer (1997), Ljung (1997), Ooi (1997), Young and Bloothoof (1997), Biber, Conrad and Reppen (1998), Kennedy (1998), Oakes (1998), Kirk (2000), Mair and Hundt (2000), and others.

The volume of reference shows that the importance of language corpora is already established in the study of language as well as in the development of many sophisticated language processing tools. Many path-

breaking researches are undertaken and many pioneering computer-based systems are developed in English and other languages using language corpora. However, the trend of using corpora in language research has taken a large amount of time to capture the fancy of Indian linguists and computer experts. Any way, we have eventually realised its importance in a multilingual country like India, and as result, the MIT, Govt. of India has initiated some works related to corpora development in all major Indian languages (Murthy and Despande 1998).

In this paper we present a brief outline how various types of language corpora can be developed for Indian languages. We also address how these corpora can be utilised in developing various language processing tools and systems for Indian languages as well as how corpora can be used in various domains of language research. The perspectives and possibilities for Indian languages can have equal relevance and importance to other languages used in South Asian countries (Bangladesh, Srilanka, Pakistan, Maldives, Nepal, and Bhutan). In section 2, we discuss briefly about the present availability of corpora in Indian languages. In section 3, we propose a scheme for developing various types of corpora for Indian languages. In Section 4, we try to show how corpora can be used for technology development and for designing various language processing tools. In section 5, we focus on multi-purpose use of corpora in various linguistic research and studies in Indian languages.

2. WHAT DO WE HAVE?

Study of Indian languages with close reference to corpus is a long due, though substantial works related to Indian English are done by some scholars. The works of Kachru (1961, 1965) are related with some creative writings of Indian English, while the works of Desai (1974) and Nihalani et al. (1979) are based on selected or 'available' samples of Indian English. We must appreciate their effort for analysis of Indian English and give them full credit for their effort. But at the same time we can argue that their works have no relation to any Indian 'national' languages used through out the subcontinent. Therefore, there is no gain in saying that a comprehensive description is needed based on corpora in Indian languages.

The first corpus in Indian languages is probably the *Kolhapur Corpus of Indian English (KCIE)* developed by Shastri (1988) and his colleagues at Shivaji University, Kolhapur, which eventually has become instrumental in creating an urge for generation of corpora for research and development in Indian languages. Their corpus consisted approximately one million words of Indian English drawn from materials published in the year 1978. Following the format used to develop the *Lancaster-Oslo-Bergen (LOB)*, and the *Brown Corpus*, the texts for *the KCIE* were selected from 15 subject categories. The data was manually input in the computer in a simple ISCI mode so that the data can be easily retrieved and accessed by the end-users. This corpus is supposed to be included as a representative sample of Indian English in the *International Corpus of English (ICE)*.

However, more than a decade have passed before we fully realised the importance and need for corpora in Indian languages. By this time, works on language corpora in Europe and America have gone a long way, and we are exposed to many new corpus-based language processing tools (lexical collocater, concordancer, local word grouper, text annotator, word tagger, sentence parser, lemmatiser, spell-checker, etc.) which have not only made research on language simple, but also have contributed handsomely on retrieving many new and relevant information form corpora for language teaching and education. This technological growth has perhaps helped us to initiate projects for corpora development in Indian languages.

In the decade of 90, initiatives are taken for corpora development in Indian languages under the patronage of the Department of Electronics (DOE), Govt. of India, in which the first author was involved. Under the *Technology Development for Indian Languages (TDIL)* programme it was decided that machine readable corpora of texts of nearly 10 million words (later modified for 3 million words) in all national languages would be developed. Software would be developed simultaneously for grammatical tagging of words, word count, frequency count, spell checkers, etc. Indian Institute of Technology, Kanpur was entrusted to develop tools for language processing and machine-aided translation from English to Indian languages (Murthy and Despande 1998: 3). We can have an idea from the following table (Table 1) how the Corpora Project was initiated and for which Indian languages, and how long it was able to continue with what results. Details of data sampling and corpus development can be found in Dash and Chaudhuri (2000).

Part	Languages	Impl. Agency	Started	Closed	Output
I	English, Hindi Punjabi	Indian Institute of Technology, New Delhi	1991	1995	3 million word corpus for each language of the group
II	Tamil , Telugu, Kannada Malayalam	Central Institute of Indian Languages, Mysore, Karnataka	1991	1995	3 million word corpus for each language of the group
III	Marathi, Gujarati	Deccan College, Pune, Maharsatra	1991	1995	3 million word corpus for each language of the group
IV	Oriya, Bangla, Assamese	Indian Institute of Applied Language Sciences, Bhubaneswar, Orissa.	1991	1995	3 million word corpus for each language of the group
V	Sanskrit	Sampurnananda Sanskrit University, Varanasi, UP	1991	1995	3 million word corpus for Sanskrit
VI	Urdu, Sindhi, Kashmiri	Aligarh Muslim University Aligarh, Uttar Pradesh	1992	1995	3 million word corpus for each language of the group

Table 1: Development of language corpora in Indian languages

However, after the completion of the project the DOE stopped the project from further continuation realising that it needs more longer investigation, and larger investment. Moreover, there was a need to change the attitude of some Indian scholars and linguistics experts who, though were not strongly against the development of language corpora, were highly sceptical about the usefulness and application of language corpora and the prospects of corpus linguistics in India. Despite their apathetic attitude a few computer scientists and linguists decided to stick to corpus linguistics, and started developing corpora and using them in various technology developments and linguistic studies in Indian languages. Details are available in *Vishwabharat*, a newsletter of the MIT. Govt. of India.

3. WHAT DO WE NEED ?

Electronically-made corpora are new things, and we are yet to reach a consensus to what counts as corpus and how it should be classified. The discipline of corpus linguistics is developing rapidly, and norms and hypothesis are under frequent revision. Therefore, any classification scheme of corpora is bound to be tentative having chances for future modification. However, we try to classify corpora in a very general way focusing on the types of need in present Indian context. They can contain written or spoken texts, new or old texts, monolingual, bilingual or multilingual texts which can be obtained from whole books, newspapers, journals, from speeches, and even from the extracts of discourse of varying length. The kind of texts included, and the combination of various text types may vary with regard to the types of corpora.

Our first need is to develop monolingual **general corpora** for each of the all major Indian languages. They will contain data comprising texts belonging to all subject domains, text and genre types, subject fields, and register variations. A good example of general corpora can be the *British National Corpus (BNC)* which contains data obtained from all walks of life. Such corpora can be fruitfully used as **monitor corpora** for various linguistic and non-linguistic studies in each language. Provision for their constant growth will reflect the changes occurring in the language as well as in the society. As a result, rarely used linguistic items will become large in number while recurrently used linguistic items will acquire archival value. Gradually, over time, they will achieve a diachronic dimension representing data obtained from a wider range of time. Such corpora will help us to identify new words and phrases, to locate newly coined technical terms, to know actual date of coming of new words and terms, to track variation in usage of lexical items and phrases, to observe changes in meaning of words, to follow changes in sentence structures, etc. All these information can only be found if we have good general corpora properly sampled and methodically designed. Neither the intuition of native-language users, nor the assumption of linguistic experts, nor even the speculation of pragmatic scholars can probably challenge the contribution of an empirical data-base in the form of corpora.

We can have annotated versions of these general corpora which will contain specific codes for both extralinguistic (e.g., text-type, year of publication, name, age, sex of the author(s), domain, source and register of texts, etc.) as well as intralinguistic (e.g., analytical marks, parts-of-speech codes, lexical category marks, grammatical category codes, sentence type codes, semantic codes, anaphoric codes, etc.) information. Although general corpora have been and are of considerable use in language study, the utility of general corpora is considerably increased by the provision of annotation (McEnery and Wilson 1996: 24). Annotated corpora will be used for theoretical linguistic analysis (e.g., study of language and gender, language change, semantic change, polysemy study, ambiguity study, etc.) as well as for designing sophisticated language processing tools (e.g., morphological processors, sentence parsers, information retrieval systems, machine translation systems, etc.) for Indian languages. Recently, Electronic Research and Development Centre (ERDC), Noida, has tagged a 3 million Hindi corpus for various grammatical categories such as verbs, nouns, adjectives, etc. The tagger tags the grammatical information to a particular word making use of their immediate context. The product can be integrated with translation packages, language learning packages, Hindi spell checker and grammar checker software (Vikas et al. 2001:32). This is high time for us to start annotating corpora of other Indian languages, available to us, in the same line proposed by ERDC.

Study on Indian speech so far has been confined mainly to aspects of spoken English (Bansal 1969) and methodological considerations (Kachru 1961). But we need is to develop **speech corpora** for all Indian languages, including dialects, if possible. This will not only help us to know the number of speech varieties used in the country but also the number of dialects used across the country. Such corpora will contain data from informal talks (e.g., gossip, banter, chatting, etc.), from impromptu conversations (e.g., quarrels, bargain, road-side talks, courtship babbles, hawkers' cry, etc.) as well as from natural dialogues with no media involvement. Such data is the most important and reliable sample of language use where any kind of media interference will definitely distort the actual image of the language. Such data is capable of reflecting the core of a language revealing most of its characteristic patterns in some way or other that no other language variety can. The ERDC, New Delhi has developed a speech database for different utterances of speech samples by large number of speakers in variety of environments and phonetic context which can be used as very basic and essential knowledge source for any research and development activity in speech technology (Vikas et al. 2001:33). Similar works are also due for other Indian languages. If required, we can transcribe speech corpora into

spoken corpora where speech data will be represented in written form with proper and uniform pattern of transcription. They can also be annotated with specific notations for standard phonetic transcription following the systems employed for transcription of speech data in English, with necessary modifications. Such phonetically transcribed corpora will be useful for those researchers who lack technological tools and expertise for laboratory analysis of recorded speech (McEnery and Wilson 1996: 26).

We should also seriously consider for developing **comparable corpora** among those Indian languages which have genealogical as well as typological similarities (e.g., Bangla-Oriya-Assamese, Telugu-Tamil-Kannada-Malayalam, Hindi-Urdu-Kashmiri, etc.). The *Comparable Reference Corpora* (of the official languages of the European Union) can be a good model for us. Such corpora will contain the collection of 'similar' texts in more than two languages or language varieties. They will consist of equal number of texts in each language, and follow the same composition pattern to be used for all kind of comparative studies among the genetically related languages. General corpora designed for each Indian language can be used for this purpose if they are designed and developed in uniform manner following similar sampling procedures, genre types, text categories, and register variations. In a multilingual country like India we can easily visualise multipurpose use of such corpora. They are indispensable resources for developing bilingual/multilingual lexicons, and bilingual/trilingual dictionaries (Landau 2001: 273-342). Moreover, they will help us to initiate cross-language research, to develop bilingual/trilingual grammars, and to enable cross-language teaching and education.

We also need to develop **parallel (aligned) corpora** (Hind-Bangla, Telugu-Tamil, Hindi-Urdu, etc.) for Indian languages. The texts of these corpora can be aligned paragraph by paragraph, sentence by sentence, phrase by phrase, or even word by word. From the study of aligned corpora we can gather information which can be utilised for devising tools to aid to automatic translation among Indian languages or between English and Indian languages. Most of the probabilistic machine translation systems are trained on such corpora (Botley, McEnery and Wilson 2000). These corpora can be best used in multilingual countries like India for communication, cross-linguistic research, cross-language teaching, etc. However, we should not ignore the need for the development of parallel corpora between English and Indian languages (English-Hindi, English-Bangla, etc.) which will be an invaluable asset for translation of domain specific information (e.g., railway information, weather reports, agricultural information, information for tour and travels, translation of government notices and circulars, medical information, etc.), and for localisation of both linguistic and non-linguistic information across the country.

There is also a need for designing **reference corpora** for all major Indian languages. Such corpora will provide comprehensive information about various aspects of Indian languages. They will be made larger enough to represent all relevant varieties of language and their vocabulary so that they can be fruitfully used as source for writing grammars, dictionaries, thesauruses, language teaching and reference materials. They will include both spoken and written, formal and informal languages representing various social and situational registers. They can be used as 'benchmark' for lexicons, for the performance of generic tools, and for specific language technology applications. In due course, with the growing influence of internal criteria, reference corpora will be used in order to measure the deviance of general corpora. We can consider *the Bank of English* as an example of reference corpus which contains nearly 175 million words, soon to top 200 million (Sinclair 1996: 105). This corpus is accepted as the base for future development of corpus-based works in Europe, USA, and elsewhere.

Finally, we can think of developing **special corpora** (e.g., corpora of dialects, corpora of language used by children, corpora of language used by linguistically impaired people, corpora of woman language, corpora of slang and criminal codes, corpora of language used by non-native speakers, corpora of very specialised areas of communication like auctioneer's patter, medical talks, gamblers' terms, etc.) which will be assembled for specific purposes. They will vary in size and composition according to the purpose of their use. Though they will not directly contribute to the description of a language because they contain a high proportion of unusual features, their relevance in the whole scope of corpus linguistics will be of significant importance because of their uniqueness in representation of particular domain or segment of the society. At certain times their origin may not be so reliable because they contain data collected from the people who are not 'normal' or who belong to a particular trade or profession. Usually, these corpora are not balanced and if they are used for other purposes they will surely give a distorted view of a language. However, the main advantage of such corpora is that their texts are selected in such a way that the phenomena one is looking for can occur much more frequently in them than in balanced general corpora. All special corpora which are enriched in this way can be used as reliable data-bases for object-oriented studies and target-specific observation.

4. CORPORA IN TECHNOLOGY DEVELOPMENT IN INDIAN LANGUAGES

Importance of speech corpora in the research of speech technology is long acknowledged even before the advent of corpus linguistics. Any development of speech processing or speech recognising systems heavily relies upon speech corpora for extracting relevant information and data as input, as well as for the verification of the outputs.

In **language pathology** specialised speech corpora are used for analysing linguistic disorders. For this, we need to develop an accurate set of abnormal speech data in Indian languages to test and verify what may be wrong with the language processing system of the linguistically disabled people. Some works on this line are done in English at Pennsylvania State University, USA and elsewhere (McEnery and Wilson 1996: 112). The *CHILDES Database* which contains a large amount of language data collected from both linguistically impaired and normal children has been analysed empirically to understand the underlying problems in the area. Moreover, the data-base is used to study the pattern of linguistic impairments among the children, and to explore what might be the factors behind their linguistic impairments. The findings have made valuable contribution in designing expert systems for repairing linguistic inabilities among impaired children. Among Indian languages a very limited work has been done in the area of linguistic impairment which lacks both quantified representative descriptions and proper methods for any generalised conclusion. We need to initiate a sincere attempt to develop specialised speech corpora of such type which can be used in understanding the actual problem in the area, as well as for developing expert systems for helping the linguistically impaired people in India.

All relevant researches in **computational linguistics** (CL) depend heavily on the data-bases and findings from corpora. Almost all software and systems of CL and natural language processing (NLP) use language data from corpora to build various computational tools like frequency counting systems, item-search engines, text summarisation systems, annotation devices, information retrieval systems, automatic translation systems, question-answering systems, etc. With the help of large language corpora these systems can perform some useful tasks which can help us in various ways in the field of language technology and processing. For us, a majority of such systems are still due because of the lack of corpora in Indian languages.

We cannot ignore the value of corpora in **word sense disambiguation** (WSD). It is agreed that context of words can provide much more information that might not be available from words isolated from their contexts of occurrence. From corpora only, we can obtain all possible contextual information of words used in a piece of text. Such contextual information can perform an important role in WSD as well as in their actual sense extraction (ASE). Long before the introduction of corpora in WSD, it is claimed that the meaning of words can be best understood by the company they keep (Firth 1957: 21). Application of corpora in WSD and ASE has enabled researchers to become more empirical and objective in approach towards exploring multisemanticity of lexical items and obtaining actual sense from the score of multiple meanings. This has made the study of semantics more challenging both in theoretical and applied levels. We can use corpora of Indian languages for developing WSD and ASE tools for Indian languages.

Bilingual aligned corpora developed in Indian languages can be useful for developing **machine translation** (MT) systems which generally use bilingual corpora as their text inputs as well as their test-beds. Availability of bilingual corpora has made significant contributions to the actual capabilities of these systems. Major domain specific bilingual aligned corpora can enable the present MT systems to adopt a self-organising approach to supplement the traditional knowledge-based approaches. Bilingual corpora along with knowledge-bases for distance-minimisation between source and target language are in regular use for domain specific machine translation technology within controlled languages where all syntactic and lexical/terminological ambiguities are suppressed beforehand (Teubert 2000: 10). Such systems not only minimise the distance of mutual intelligibility but also enhance the translatability between the two languages which are used for designing bilingual corpora. If we really want to develop automatic machine translation systems among Indian languages then we cannot ignore the need for bilingual corpora as well as their usefulness in the development of these systems.

All general corpora developed in Indian languages can be effectively used in **computer-assisted language teaching** in each Indian language. Corpora-based approaches are more effective in language teaching and learning than traditional lecturer-based methods because these approaches are capable in furnishing the actual evidence of language use which is invariably missing in lecture-based methods. It proved that the use of corpora and computer in language teaching produces much better results than intuition-controlled classroom teaching (McEnery and Wilson 1996: 105). What is more encouraging is that corpora and computer-taught students perform far better than human-taught students (McEnery, Baker and Wilson 1995: 259-274, Barlow 2000:106-115, Salkie 2000:148-156). Corpora are beneficial for the language learners to understand how various properties of language are used actually in texts. By citing examples of actual patterns of use of various linguistic items from corpora we can equip and inform them for better understanding and producing language appropriately across different contexts and situations (Biber, Conrad and Reppen 1998:82). Language corpora can also be useful for teaching them about different lessons, for answering various queries, and for responding to their subject-specific questions. In countries like India development of such systems can help in achieving good results in language teaching as well as in enhancing linguistic skills of language learners. Both primary students as well as newly literate people can find language teaching more interesting and useful if linguistic information and examples are obtained from the corpora of their languages. Thus, corpora-based approach will enable language learners to:

- (i) identify properly the actual categories of lexical items used in texts
- (ii) recognise the proper contextual uses of various lexical items

- (iii) understand the appropriate use of scientific/technical words and terms depending on the context or topic of texts
- (iv) know how various types of sentence are actually used in different text types
- (v) understand how idioms, phrases, proverbs etc. are used in texts
- (vi) to realise how context can change word meanings referred in standard dictionaries, etc.

Finally, general corpora in Indian languages can be used as aids for different kinds of **text preparation** (e.g., writing articles, books, revising written texts, etc.) after they are properly processed and categorised. Various corpus-based tools like word processors, spell checkers, etc. can also access general corpora for their tasks as well as for the modification of the systems. Word processors in Indian languages can use data from corpora to prepare and edit texts while spell-checkers in various Indian languages can use data from corpora to correct errors in spelling, and to provide spelling alternatives (Winograd 1983: 26). The commercial importance of such tools in a multilingual country like India is highly lucrative.

We can sum up how corpora in Indian languages can be actively utilised in various domains of technological development in Indian languages:

A. Knowledge resources

- Development of multilingual libraries
- Development of monolingual, bilingual and multilingual dictionaries (both printed and electronic version)
- Development of machine readable dictionaries (MRDs)
- Development of multilingual lexical resources

B. Language tools

- Word processing systems
- Spell checker systems
- Text editing systems
- Morphological analysing systems
- Sentence parsing systems

C. Translation support systems

- Language resource access systems
- Machine translation systems
- Multilingual information access systems
- Cross language information retrieval systems

D. Human-Machine interface systems

- OCR systems
- Voice recognition systems
- Text-to-speech systems
- Web-based learning systems

Furthermore, various types of frequency count systems, concordancing systems, parsing systems, text annotation systems, etc. can be designed and used for processing Indian language corpora and for retrieving relevant information from them. Information acquired by the application of these systems can be used for designing coursebooks for language learners, building OCR systems, developing spell-checkers, designing tools for machine translation, creating lexical data-bases, developing dictionaries, etc. in all Indian languages. The

products can be of useful for language learners, researchers, writers, teachers, students, scholars, publishers, and all.

5. CORPORA IN RESEARCH AND EDUCATION IN INDIAN LANGUAGES

Language corpora are important source of data for a number of areas within the wide scope of language research and education. Methodical analysis of language corpora can provide important complementary perspective to traditional linguistic descriptions (Biber 1996: 173). We can retrieve appropriate information from corpora to employ them in various linguistic studies such as lexicography design, writing grammars, semantic study of words, pragmatic analysis of texts, sociolinguistic study, discourse analysis, etc. (Leech and Fligelstone 1992: 129). In the following sections we present brief discussions on how corpora can be used in various fields of linguistic research and education.

Close comparative analysis of languages used in different states of India can help us to know the **cultural similarities** and **differences** existing among the people of different language communities. For example, a comparative study on the vocabulary of Bangla and Hindi corpora can show which type of lexical items and phrases are mostly used in them, and what kind of social life is reflected by the use of these vocabularies. A simple comparative analysis between Hindi *yAr* [iar] and Bangla *dAdA* [dada] shows many interesting aspects on the use of such addressing terms in both the language communities. Literally, *yAr* means "friend" in Hindi, while *dAdA* means "elder brother" in Bangla. The notable point is that both the terms are randomly used by the users in almost all register variations or text types. Both in Hindi and Bangla speaking community, everybody irrespective of any gender, age, class, and race can be addressed as *yAr* or *dAdA* in all informal situations. However, a finer line of distinction can reveal that *yAr* is a **social term** literally used for addressing the people who belong to same social class, age, profession or fields, while *dAdA* is **family term** which is normally used for addressing elder brothers or brethren belonging to family or relatives. We can guess that the use of *yAr* can probably hint for keeping everybody (both known and unknown) within the circle of social proximity, while use of *dAdA* closes the circle a little further to bring it at family level.

Research on **grammar** and **syntax** of a language is usually carried out based on the investigator's intuition about the language rather than on the performance of language users. "Every (formal) grammar is initially written on the basis of intuitive data. By confronting the grammar with the unrestricted corpus data it can be tested on its correctness and its completeness" (Aarts 1991: 45). Availability of corpora can make it possible to study Indian languages as they are actually used by the people. Thus, Indian language corpora can be used for grammatical and syntactic studies in Indian languages as they can give us scopes for:

- (i) evaluating existing grammars of Indian languages
- (ii) testing hypotheses derived from various grammatical theories
- (iii) testing formal grammatical descriptions devised on the basis of investigators' introspection
- (iv) finding out how far the standard grammars can account for the data used in corpora, and
- (v) modifying grammars to account for the situations which are either missing or found wrong

In recent times, large corpora are used for many small-scale studies on grammar and syntax (Oostdijk and Haan 1994, Mindt 1995) in English. Such kind of study in Indian languages will be highly beneficial to the language learners as well as Indian language technology developers. Besides, we can develop corpora-based grammars in Indian languages which are able to represent the state-of-the-art of the languages as they will faithfully reflect how Indian languages are being used in reality. Such grammars (e.g., *COBUILD Grammar of Modern English*) are good for both language teaching as well as parsing of sentences. Corpora-based **core grammars** are fine tools for cross-linguistic research and education as they provide opportunities for

interlingual studies, information interchange, translation, etc. In a multilingual country like India such grammars can probably enhance national integrity through interlingual communication^[1] and information interchange.

If large diachronic corpora in Indian language are developed they can be used extensively in the study on **historical development** of Indian languages. Diachronic Indian corpora can be good resources for the study of evolution of Indian languages through time. Fruitful diachronic language research can be conducted faster and more efficiently with such corpora representing languages at different points in history. The data of historical corpora can be extended by the discovery of previously unknown documents in the form of manuscripts or books. Such corpora covering wider range of genres, regional varieties, and sociolinguistic variables (e.g., gender, age, education, social class, etc.) can be used for the purpose of both synchronic and diachronic research within or across languages in India.

In **applied linguistics** corpora in Indian languages can be used as reliable data-bases for language teaching, language acquisition, as well as for improving linguistic skills of language learners. Apart from being source of empirical data-base, corpora can be used to evaluate existing language teaching materials also. It is argued that there are considerable differences between what textbooks are teaching and how native speakers actually use language (Botley, McEnery and Wilson 2000). The general opinion is that intuition based teaching materials can be misleading since they contain intentionally invented examples which generally overlook important aspects of usage, or foreground less frequent stylistic choices at the expense of more common ones. But teaching materials should be explicitly empirical including examples and descriptions from corpora or other sources of real life language data so that more common choices of usage are given more attention than those which are less common. All these have inspired Barlow (1996: 2) to argue that:

"Corpora can reveal not only the range of patterns of a language that the learner must assimilate, but also their frequency, which is an important factor in materials development and syllabus design. Moreover, the use of corpora has the potential to radically alter the field of second language learning, and perhaps linguistics as well".

Corpora in Indian languages will be valuable sources for comparison and description of **language variations** present in India. Following some sampling procedures such corpora can be designed to maximise the degree of comparability among the related Indian languages. For example, both Bangla and Hindi corpora can contain texts representing similar genres, sample size, and year of publication to enhance synchronic comparative study. Thus, they can be used to study how language varies according to different texts, genres, domains, periods, regions, speakers, writers, contexts, etc. The 'variants' can either be different parts of one corpus (e.g., science fiction texts vs. romantic fiction texts), or similar parts of different corpora (e.g., science fiction texts in Bangla corpora vs. science fiction texts in Hindi corpora). Such comparable corpora can also be used as test-beds for theory making as well as for studying language variations. Intensive comparative studies on lexis and grammars of Hindi, Bangla, Telugu and other corpora can probably substantiate the Common Core Hypothesis^[2] (Quirk et al. 1985) to enhance linguistic integrity among Indian languages.

Corpora developed in the dialects of Indian languages can be used for the study **dialects** of Indian languages. Dialect study is mostly empirical which usually relies on experiment and controlled data-sampling rather than on elaborate corpora. Such studies tend to focus on vocabulary and pronunciation neglecting other aspects of language such as morphemes, sentences, meanings, etc. which we can study using dialect corpora. If dialect corpora are accurately sampled to be properly representative then we can make both quantitative as well as qualitative conclusions about the target population as a whole. A good model for our work can be the study

where *the Helsinki Corpus of English Dialects*, *the Northern Ireland Transcribed Corpus of Speech*, *the LOB Corpus*, and *the London-Lund Corpus* which are used to examine the degree of similarity and difference of dialects as compared with the standard varieties of English (McEnery and Wilson 1996: 110).

For a long time in language study empirical data is utilised in **lexicography**. Samuel Johnson used examples from literature to illustrate his dictionary (1755), while the *Oxford English Dictionary* used citation slips to study and illustrate the usage of words in English. Today, corpora are used for dictionary preparation as they play important role in dictionary building. The lexicographers can use corpora to:

- gather evidence that can either supplement or refute their intuitions
- find new words entering the language
- identify how existing words changing their meanings
- recognise how existing words balance their use according to genre, texts, etc.
- call up all the examples of particular words to substantiate their contextual variations
- revise existing dictionaries
- present more complete and precise definitions of various linguistic items
- give updated information about any change or loss of any word in the language
- arrange examples extracted from corpora into more meaningful groups for analysis
- classify individual words according to different research parameters
- isolate word combinations to explore the existence of any inherent mutual relationship which warrants their co-occurrence, etc.
- treat phrases and collocations more systematically because phrases and collocations can provide important clues to specific word-sense, and
- to tie up usages of particular words or phrases as being typical to particular regional varieties, genres, and so on, by examining corpora rich with textual and extra-textual information (e.g., regional variety, author, date, sex, genre, part-of-speech tags, etc.).

Besides the works mentioned above Indian language corpora can also be used in lexical quantification, and in subdivision of areas across the varieties of language in which words are used. In Indian perspective we can think of using corpora for developing both monolingual and bilingual dictionaries as well as dictionaries of technical and scientific terms, spellings, polysyms, homonyms, paronymns, acronyms and abbreviations, and many others. Such dictionaries will definitely be much more advanced, informative and realistic than the dictionaries available in the market. In 1995, we have come across the publication of four major English dictionaries which are entirely based on corpora^[3]. These widely acclaimed dictionaries contain information about lexical sub-categorisation, contextual restriction, polysemous connotation of the entries besides their phonetic, morphological, semantic, and syntactic information which are retrieved from corpora.

Information can be obtained from corpora for laboratory experiments in **psycholinguistics**. Frequency of lexical items is an important consideration in a number of cognitive processes including wordform recognition and understanding in natural languages (Biber 1996). Corpora can provide reliable information about the frequency of use of different lexical items in a language for making hypotheses about their method of processing in mind. Generally, speech corpora are recommended for the examination of occurrence of speech errors in natural conversation. After studying *the London-Lund Speech Corpus* for speech errors in natural conversation scholars have been able to classify and count the frequencies of different error types in speech. Moreover, they have been able to provide some estimate on the general frequency of these in relation to

speakers' overall linguistic output (Garnham et al. 1981). Such kind of studies focus on the need for the development of speech corpora in Indian languages if we are really interested to know what kinds of error generally occur in our everyday speech, and what kind of steps we can take up to overcome such errors.

Corpora can contribute towards the establishment of an objective approach to the **semantic study** of linguistic items and text segments. It is argued that actual meaning of lexical items (or any text segment) can only be derived from the contexts in which they occur. Also, the meaning of lexical items or text segments comprises the history of their previous occurrences which include everything that has been said there about their meaning and about the meaning of the parts they are composed of. In semantics, the meaning of terms are quite frequently described by reference to the linguist's own intuitions. However, Mindt (1991) has demonstrated how corpora can be used to provide objective criteria for assigning meaning to the linguistic terms. He has shown that semantic distinctions are associated in texts with morphological, syntactic, prosodic, idiomatic or/and similar such characteristically observable contexts. Therefore, compounds, multi-word units, collocations, and set phrases deserve contextual information for their proper semantic analysis, and understanding. We can argue that an empirical objective interpretation for particular semantic distinctions can be arrived at by considering the environments of linguistic entities. Moreover, we can extract the finer shades of meaning (denotative, connotative, conceptual, pragmatic meaning, etc.) from the contexts, condense and paraphrase them into texts that describe meaning of lexical items or text segments (Teubert 2000).

The use of corpora in the study of **polysemy** (multisemanticity of single lexical item) has helped semanticists to establish more firmly the notion of semantic gradience in the field of lexical semantics (Cruse 1986). Examples obtained from the dictionaries compiled manually, and from the dictionaries compiled from corpora, have shown that the number of sense distinctions of a particular lexical item that show up in corpora far exceeds the number of sense distinctions that are provided in dictionaries (Fillmore and Atkins 2000: 91-110). It is argued that polysemy of words can be rightly understood if we include corpora-based lexicography along with other relevant properties of lexical items. Such approach might be useful for capturing multiple semantic senses of words in monolingual corpora, as well as for sense identification and matching for translation-equivalents among bilingual corpora.

Majority of research in **sociolinguistics** usually rely on the collection of research-specific language data which are rarely put to any quantitative study or sampling. Moreover, these studies are usually concerned with sets of lexical items related to the area of language and gender. Corpora with multi-dimensional sociolinguistic information are highly useful for various sociolinguistic researches because they can provide representative samples of naturalistic data which can be used in both kinds of investigation. We can use corpora of any Indian language (with various sociolinguistic information such as users' age, sex, social class, education, domicile, economic status, etc.) to examine if there is any masculine bias in the language, to find the ratio of use of feminine terms in respect to masculine items, and to evaluate the relevance of use of feminine/masculine terms in texts. The findings may provide subtle clues for determining status of women in the society where they are supposed to have equal status like their counterparts. Recently, *the Corpus of London Teenagers* is used to study variations in the verbal disputes of adolescent females (Stenstrom and Hasund 1996) while *the British National Corpus* is used to study various sociolinguistic issues (Rayson and Hodges (1997) in English.

For studying **speech** varieties used in India the speech corpora of Indian languages are of great value. These speech corpora can be used as reliable data-bases for studying speech patterns of the speakers coming from different walks of life. Such study can produce results which can virtually differ from the results obtained from written corpora. These speech corpora can also be used for teaching computers to synthesise, recognise,

understand, process, and produce speeches spoken in the country. Besides, these speech corpora can be utilised for developing various speech technology in Indian languages. The importance of speech corpora can be summarised as follows:

- (i) Speech corpora can provide a broad representative samples of speech extending over a wide selection of variables (e.g., speaker's gender, age, class, etc.),
- (ii) Speech corpora can capture both the generalisations about spoken language as well as the variations within a given spoken language,
- (iii) Speech corpora can supply us reliable samples of naturalistic speeches rather than speeches developed under artificial conditions,
- (iv) The findings from speech corpora are more likely to reflect language as it is spoken in real life situations,
- (v) Speech corpora with phonetic and prosodic annotation are far more suitable for carrying out large scale quantitative analysis than fresh raw data,
- (vi) The study of acoustic and phonetic aspects of actual speech data is important in the areas of telecommunication, voice-mail systems, etc.

Corpora in Indian languages with genre and register variation information can open up many new possibilities of research in **stylistics**. We can access texts with specific features of certain text types in place of general varieties of language. Because of their comparability and referentiality, corpora are good sources for investigating different text and genre types within monolingual environment for authorship attribution, and for investigating changes of styles. In order to define an author's particular style, we can examine the degree to which the author leans towards different ways of putting things (technical vs. non-technical vocabulary, long vs. short sentences, formal vs. informal manner of narration, etc.). Such tasks require comparison of the author's own works, as well as comparison with works of other authors, with the norms and variety of the language as a whole. We can argue that Indian corpora can be reliable data-bases for the study on text types, as well as for authorship attribution following particular styles of certain authors.

We can think of using Indian language corpora in **discourse analysis** though it is difficult to use written corpora for such research because the samples of texts used in written corpora are somewhat removed from their social and temporal contexts which are of paramount importance in discourse analysis. To make up this loss we can encode the corpora with relevant social and demographic clues (e.g., gender, class, region, religion, age, etc.) to infer some information about their contexts. Using the *London-Lund Speech Corpus*, Stenstrom and Andersen (1996), Andersen (1997) and others have carried out some studies to understand how conversation works with respect to lexical items and phrases that have conversational functions. We can expect that corpora in Indian languages will be annotated with relevant information to trigger corpora-based discourse research in India. Moreover, we can develop conversational corpora in Indian languages as well as accumulate recorded data of various social/geographical range to provide an incentive to the study on discourse.

Finally, corpora in Indian languages after being properly processed and categorised can be used as aids for various kinds of text preparation such as writing articles, books, papers, etc. Various computer tools like word processors, spell checkers etc. can use data from corpora for preparing and editing texts, for correcting errors in spelling, for providing spelling alternatives etc. (Winograd 1983: 26). Such tools have proved to have strong commercial viability in India and abroad.

In this section we have tried to understand how corpora can be best used in language research and education in a multilingual country like India, and how they can provide good empirical support for studies in

different branches of linguistics. After nearly five decades of use of corpora in different sub-domains of linguistics, psychology, statistics, and information technology, it is now almost clear to us that the application of corpora is not confined within teaching and research of language. We can think of many more new applications of corpora. In a sample study, we have found that the English corpora have been used for about five hundred different types of research and education covering almost all domains of linguistics^[4]. The multifunctionality of corpora is, perhaps best perceived by Svartvik (1986: 9) who has visualised that corpora can be used in:

"lexicography, lexicology, syntax, semantics, word-formation, parsing, question-answer synthesis, software development, spelling checkers, speech synthesis and recognition, text-to-speech conversion, pragmatics, text linguistics, language teaching and learning, stylistics, machine translation, child language, psycholinguistics, sociolinguistics, theoretical linguistics, corpus clones in other languages such as Arabic and Spanish - well, even language and sex".

The scope is further expanded by Atkins et al.(1992), Leech and Fligelstone (1992), McEnery and Wilson (1996), Rundell (1996), Barlow (1996), Thomas and Short (1996), Biber (1996), Biber, Conrad and Reppen (1998), Teubert (2000), and others. Anyway, we can sum up the use corpora in the following way:

- (i) Corpora will become natural and indispensable resource in the field of general language study, description, and teaching.
- (ii) Corpora will become the most reliable treasure-house for creation of various dictionaries and reference books (both monolingual and bilingual).
- (iii) Corpora will become indispensable in the development of various language processing tools, systems, and software.
- (iv) Corpora, for their easy availability and fast accessibility in machine-readable form, will become ready-made source for multi-purpose (mostly non-linguistic) use by the end-users.
- (v) Corpora can be customised for studying some particular area of interest.

We can guess that corpora will grow up with time and will include large sub-sections of texts classified by date, subject-matter, region, age-group, sex, etc. Thus, they will become valuable data source to the people who create them as well as to the people who are interested in text's language. We can visualise the following people will be highly interested in language corpora:

- (i) Language specialists who visualise corpora as large diluted source of data that can be used as yardstick for any kind of linguistic and non-linguistic verification and analysis,
- (ii) Language applicers who consider corpora as test-beds, composed of representative materials, and good for testing or training automatic devices.
- (iii) People who are working in NLP in Indian languages will require large amounts of language data for their research and developments,
- (iv) Lexicographers who need 'balanced' corpora reflecting a wide range of text-types, and
- (v) Some enthusiasts working in various areas of linguistics who happen to be interested in corpus-study

6. CONCLUSION

For last 40 years or so corpora are considered as basic resources for language analysis and research. This reflects both ideological (a major shift towards empirical study of language from intuitive or rationalistic assumptions) and technological (impressive processing power and massive storage ability of computers) change in the area of language research. This change of attitude is probably caused due to the introduction of computer

and corpora in linguistic research which, as a result, have paved out many new applications of language (and linguistics) in the fields of communication and information exchange. Moreover, the empirical approach to language study has been identified to be more reliable and authentic than rationalistic (based on intuition) approaches.

The use corpora in Indian languages for various technological developments as well as for various linguistic studies in Indian languages can open up many new avenues for us. These corpora can be useful for producing many sophisticated automatic tools and systems, besides being good resources for language description and theory making. This is high time for us to turn our attention towards corpora for our linguistic research and description. Otherwise, we have every chance to be non-reliable and misleading in our attempts for describing Indian languages without proper reference to our empirical language data-bases.

ACKNOWLEDGEMENT

An earlier version of this paper was presented in the International Conference of *SCALLA2001 of Sharing Capability in Localisation and Human Language Technologies (SCiLaHLT)* funded by the European Commission, at NCST, Bangalore, India during 21-23rd November, 2001. The paper is modified to a great extent after receiving feedback from the participants of the Conference. The authors like to thank all participants for their wise comments and valuable insights.

NOTES

- [1] In fact, this was one of the reasons behind the initiation of the TDIL project of DOE, Govt. of India, in 1991 (See Murthy and Deshpande 1998)
- [2] Common Core Hypothesis (CCH) is a theory proposed by Quirk et al. (1985) which has assumed that all varieties of English in the world have certain central fundamental properties in common which may differ quantitatively rather than qualitatively.
- [3] The dictionaries are: *Longman Dictionary of Contemporary English* (3rd ed.), *Oxford Advanced Learner's Dictionary* (5th ed.), *Collins Cobuild English Dictionary* (2nd ed.), and *Cambridge International English Dictionary* (3rd ed.). For details, see Rundell (1996).
- [4] One can surf the bibliography of ICAME home-page <www.khnt.hit.uib.no/icame/manuals/icambib3.htm> compiled by Bent Altenberg of Department of English, Lund University, Sweden. It contains only a partial information on works done on English corpora. It can be assumed that similar such works are/can be done in other languages. However, no information of such work on Indian language corpora is available to us.

REFERENCES

- Aarts, J. and W. Meijs (eds.) 1984. *Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research*. Rodopi: Amsterdam-Atlanta, GA.
- Aarts, J. and W. Meijs (eds.) 1986. *Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora*. Rodopi: Amsterdam-Atlanta, GA.
- Aarts, J. 1991. "Intuition-based and Observation-based Grammars". *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, eds. by K. Aijmer and B. Altenberg, 44-62. London: Longman.
- Aijmer, K. and B. Altenberg (eds.) 1991. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman.
- Andersen, G. 1997. "They like wanna see like how we talk and all that. The use of *like* as a discourse marker in London teenage speech". *Corpus-based studies in English*, ed. by M. Ljung, 37-48. Rodopi: Amsterdam-Atlanta, GA.
- Atkins, S., J. Clear and N. Ostler. 1992. "Corpus Design Criteria." *Literary and Linguistic Computing*. 7(1): 1-16.
- Bansal, R.K. 1969. *The intelligibility of Indian English*. Monograph No. 4, CIEFL, Hyderabad.
- Barlow, M. 1996. "Corpora for Theory and Practice." *International Journal of Corpus Linguistics*. 1(1): 1-38.
- Barlow, M. 2000. "Parallel texts in language teaching". *Multilingual Corpora in Teaching and Research*, eds. by S. P. Botley, A. M. McEnery and A. Wilson. 106-115. Rodopi: Amsterdam-Atlanta, GA.
- Biber, D. 1996. "Investigating Language Use through Corpus-based Analyses of Association Patterns." *International Journal of Corpus Linguistics*. 1(2): 171-198.
- Biber, D., S. Conrad and R. Reppen 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Botley, S.P., A.M. McEnery and A. Wilson eds. 2000. *Multilingual Corpora in Teaching and Research*. Rodopi: Amsterdam -Atlanta, GA.

- Cruse, D.A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Dash, N.S., and B.B. Chaudhuri. 2000. "The process of designing a multidisciplinary monolingual sample corpus." *International Journal of Corpus Linguistics*. 5(2): 179-197.
- Desai, S.K. 1974. *Experimentation with language in Indian Writing in English (Fiction)*. Monograph of the Dept. of English, Shivaji University, Kohlapur.
- Firth, J.R. 1957. "Modes of Meaning". *Papers in Linguistics* 1934-1951. Oxford: Oxford University Press.
- Fillmore, C.J. and B.T.S. Atkins. 2000. "Describing Polysemy: The Case of 'Crawl'". *Polysemy*, eds. by Y. Ravin and C. Leacock. 91-110. Oxford: Oxford University Press.
- Garnham, A., R. Shillock, G. Brown, A. Mill and A. Cutler. 1981. "Slips of the tongue in the London-Lund corpus of spontaneous conversation." *Linguistics*. 19: 805-817.
- Garside, R., G. Leech and G. Sampson. eds. 1987. *The Computational Analysis of English: A Corpus Based Approach*. London: Longman.
- Garside, R., G. Leech and T. McEnery. eds. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Addison-Wesley Longman.
- Johansson, S. and A-B. Stenstrom. eds. 1991. *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.
- Kachru, B.B. 1961. *An analysis of some features of Indian English: A study in linguistic Method*. Unpublished doctoral dissertation. Edinburgh University, UK.
- Kachru, B.B. 1965. The Indianness in Indian English. *Word* 2: 391-410.
- Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison-Wesley Longman.
- Kirk, J.M. ed. 2000. *Corpora Galore: Analyses and Techniques in Describing English*. Rodopi: Amsterdam-Atlanta, GA.
- Knowles, G., B.J. Williams and L. Taylor. eds. 1997. *A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English Corpus*. United Kingdom: Longman Group.
- Lancashire, I., E. Carol and C.F. Meyer. eds. 1997. *Synchronic Corpus Linguistics*. Papers from the 16th International Conference on English Language Research on Computerized Corpora: ICAME 16.
- Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. (2nd Ed.) Cambridge: Cambridge University Press.
- Leech, G., and S. Fligelstone. 1992. "Computers and Corpus Analysis." *Computers and Written Texts*, ed. by C.S. Butler. 115-140. Oxford: Blackwell Publishers.
- Ljung, M. ed. 1997. *Corpus-based studies in English*. Rodopi: Amsterdam-Atlanta, GA.
- Mair, C., and M. Hundt. eds. 2000. *Corpus Linguistics and Linguistics Theory*. Rodopi: Amsterdam-Atlanta, GA.
- McEnery, T., J.P. Baker and A. Wilson. 1995. "A Statistical analysis of corpus based computer vs traditional human teaching methods of part of speech analysis." *Computer Assisted Language Learning*. 8: 259-274.
- McEnery, T., and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Mindt, D. 1991. "Syntactic evidence for semantic distinctions in English." *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, eds. by K. Aijmer, and B. Altenberg. 182-196. London: Longman.
- Mindt D. 1995. "An Empirical Grammar of the English Verb: Modal Verbs." From <http://www.engdidakt.fu-berlin.de/english/emp-grammar-mod.html>.
- Murthy, B.K. and W.R. Despande. 1998. "Language Technology in India: Past, Present, and the Future." Proceedings of the SAARC Conference on Extending the use of Multilingual and Multimedia Information Technology (EMMIT'98). Pune.
- Nihalani, P., R.K. Tongue and P. Hosali. 1979. *Indian and British English: A handbook of Usage and pronunciation*. New Dehli: Oxford University Press.
- 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 P. deHann. eds. 1994. *Corpus Based Research into Language*. Rodopi: Amsterdam-Atlanta, GA.
- Oostdijk, N., and P. deHaan. 1994. "Clause patterns in Modern British English: A corpus-based (quantitative) study." *International Computer Archive of Modern English (ICAME) Journal*. 18: 41-80.
- Percy, C., C. Meyer and I. Lancashire. eds. 1996. *Synchronic Corpus Linguistics*. Rodopi: Amsterdam-Atlanta, GA.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rayson, P., and M. Hodges. 1997. "Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus." *International Journal of Corpus Linguistics*. 2(1): 133-152.
- Rundell, M. 1996 "The Corpus of the Future and the Future of the Corpus". Talk at a special conference on *New Trends in Reference Science* at Exeter, UK (a printed hand out).

- Salkie, R. 2000. "Unlocking the power of the MEMUC." *Multilingual Corpora in Teaching and Research*, eds. by S.P. Botley, A.M. McEnery and A. Wilson. 148-156. Rodopi: Amsterdam-Atlanta, GA.
- Shastri, S.V. 1988. "The Kolhapur Corpus of Indian English and work done on its basis so far." *International Computer Archive of Modern English (ICAME) Journal*. 2:15-26.
- Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. 1996. "The Empty Lexicon". *International Journal of Corpus Linguistics*. 1(1): 99-120.
- Stenstrom, A-B. and G. Andersen. 1996. "More trends in teenage talk: A corpus-based investigation of the discourse items *cos* and *innit*." *Synchronic corpus linguistics* eds. by C. Percy, C. Meyer and I. Lancashire. 189-203. Rodopi: Amsterdam-Atlanta, GA.
- Stenstrom, A-B. and I. K. Hasund. 1996. "Girls' conflict talk: a sociolinguistic investigation of variation in the verbal disputes of adolescent females". *A Study from COLT Corpus of London teenager language*. University of Bergen. Paper presented at ICAME, Stockholm. (a hand out).
- Svartvik, J. 1986. "For Nelson Francis". *ICAME News*. No. 10: 8-9.
- Svartvik, J. ed. 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter.
- Teubert, W. 2000. "Corpus Linguistics-A Partisan view." *International Journal of Corpus Linguistics*. 4(1):1-16.
- Thomas, J., and M. Short. eds. 1996. *Using corpora for language research. Studies in the honour of Geoffrey Leech*. London and New York: Addison Wesley Longman.
- Vikas, O., P.K. Chaturvedi, S. Lata, V.K. Sharma and M. Jain. 2001. *Vishwabharat* (Indian Technology Newsletter 3), Sept. 2001.
- Winograd, T. 1983. *Language as A Cognitive Process*. Vol.I. Mass.: Addison-Wesley.
- Young, S., and G. Bloothoof. eds. 1997. *Corpus-Based Methods in Language and Speech Processing*. Vol-II. Dordrecht: Kluwer Academic Publishers.
