LANGUAGE CORPORA: PRESENT INDIAN NEED

NILADRI SEKHAR DASH
Indian Statistical Institute, Kolkata
Email: niladri@isical.ac.in

ABSTRACT

Corpora have proved their value both in linguistics and language technology. Information obtained from corpora has challenged the intuitive language study, since intuitive observations are found inadequate while compared with findings from corpora. However, the value of corpora is not yet acknowledged in India, although in recent times some sporadic attempts are made for designing corpora in Indian languages. We argue here for initiating large-scale projects to develop corpora of various types in Indian languages not only to contribute in research of language technology, but also to provide reliable language resources for the benefit of people of the country. We plea for the generation of specific types of corpus required for designing tools and systems for language technology linguistics research, and education.

1. INTRODUCTION

Utilisation of language corpora in Language technology and linguistics research is an established truth. However, this is yet to culminate in India. Here language technology is in its infancy, linguistic activities are in traditional mode, and language education follows centuries-old pedagogic process. All these pay no attention to language corpora. As a multilingual country, India is a linguistic giant. It preserves large number of living languages of various ethnic and linguistic communities. Due to lack of corpora, these languages suffer from the scarcity of language technological advancements. Generation of corpora could enhance language education to increase literacy rate, save endangered Indian languages from extinction, protect languages, which have lost relevance against the overwhelming aggression of English. Corpora could help these to survive in the battle of linguistic imperialism. In addition, it will supply statistically reliable information to regain their lost ground.

In this paper, we seek to draw attention to that fact that India, in comparison to other advanced countries, lags far behind in respect to corpus generation, LT development, corpus-based language research, and education. It becomes more painful when we realise that some Indian languages (e.g. Hindi, Bangla, Urdu, Telugu, Malayalam, Tamil, etc.) have much larger number of speakers than the languages of some advanced countries. We believe that designing tools and systems for language technology in Indian can be best achieved after the generation of corpora in Indian languages. Therefore, we propose for the generation of written and speech corpora in all Indian languages. This proposal is equally relevant to other languages spoken in the South East Asian countries (Bangladesh, Srilanka, Pakistan, Maldives, Nepal, and Bhutan).

2. EARLY ATTEMPTS

The first corpus in Indian language is the *Kolhapur Corpus of Indian English (KCIE)*, which is designed by Prof. S.V. Shastri and his colleagues at the Shivaji University, Kolhapur, India in 1988. It contains approximately one million words of Indian English drawn from materials published in the year 1978. This is made to serve for a comparative study among the American, the British, and the Indian English. It has ably projected the independent entity of the Indian English, which is rich with Indian vocabulary and

terminology, different with unique syntactic pattern and semantic load, and free from the shadow of the British English. It has declared the Indianness of Indian English as a post-Independence phenomenon, which has reached to a discernible stage in the last thirty years after independence. Despite some limitations and shortcomings of the corpus, the effort deserves credit and appreciation. However, it has little relevance to any Indian 'national' language used in the country. The work has failed to create an urge for generation of corpora in Indian languages.

4. PRESENT SCENARIO

The all-round growth of LT in advanced countries made the Indian experts realise the value of corpus in a multilingual country like India. This made them initiate projects for corpus development nearly a decade ago. Consequently, the Department of Electronics (DOE), Govt. of India started corpus development projects in Indian languages from the 1991 in which the present author was involved. In 1991, under the Technology Development for Indian Languages (TDIL) programme, it was decided that machine-readable corpora of nearly 10 million words would be developed within three years for all Indian national languages. Software for POS tagging, frequency count, spell-checkers, morphological processing, etc. would also be developed for Indian languages using the corpora. Indian Institute of Technology, Kanpur was entrusted to develop tools for language processing and machine-aided translation system from English to Indian languages (Murthy and Despande 1998: 3). We can have some idea from the Table 1 how the corpus generation project was initiated and worked out. Details are discussed elsewhere (Dash and Chaudhuri 2000).

Part	Language	Agency	Started	Closed	Word
I	English, Hindi, Punjabi	IIT,New Delhi	1991	1994	3 million
II	Telugu, Kannada, Tamil, Malayalam	CIIL, Mysore	1991	1994	Do
III	Marathi, Gujarati	DC, Pune	1991	1994	Do
IV	Oriya, Bangla, Assamese	IIALS, Bhubaneswar	1991	1994	Do
V	Sanskrit	SSU, Varanasi	1991	1994	Do
VI	Urdu, Sindhi, Kashmiri	AMU, Aligarh	1992	1994	Do

Table 1: Generation of text corpora in Indian languages

Agencies	Languages
Indian Institute of Technology, Kanpur	Hindi, Nepali
Indian Institute of Technology, Mumbai	Marathi, Konkani
Indian Institute of Technology, Guwahati	Assamese, Manipuri
Indian Institute of Sciences, Bangalore	Kannada, Sanskrit
Indian Statistical Institute, Kolkata	Bangla
Jawaharlal Nehru University, New Delhi	Sanskrit
University of Hyderabad, Hyderabad	Telugu
Anna University, Chennai	Tamil
MS University, Baroda	Gujarati
Utkal University, Bhubaneswar	Oriya
Thapar Institute of Engg. and Tech., Patiala	Punjabi
ER&DCI, Trivandrum	Malayalam
C-DAC, Pune, Maharastra	Urdu, Sanskrit, Kashmiri

Table 2: Agencies engaged in corpus development in Indian languages

However, by the end of 1994, when corpora of 3 million words was developed for each Indian language, the DOE declined from further continuation of the project, probably understanding that the project needs far more time and investment. Unfortunately, this negative move was applauded by some Indian experts, who were sceptical about the usefulness and application of corpus in Indian contexts. Notwithstanding this air of antipathy, some Indian scholars decided to stick to corpus generation and processing with an intention for using it in LT development and linguistics studies. Such effort is eventually acknowledged by the Ministry of Information Technology (MIT), Govt. of India, which has recently taken steps to rejuvenate the projects closed decades ago (Vikas at al. 2001). At present, in a partial estimation, the following agencies (Table 2) in India are engaged in developing corpus in Indian languages [1].

5. WHAT DOES INDIA NEED?

5.1 WRITTEN CORPUS

At present we dearly need written corpora (monolingual, bilingual, and multilingual) for all Indian languages. A general monolingual written corpus of each language would contain data from all subjects, domains, texts, genres, types, fields, and occupations. Like the *British National Corpus (BNC)* it can be used as monitor corpus (Sinclair 1991: 24-26) for various linguistic and non-linguistic studies, cross-linguistic comparisons, and all other areas of LTs. It may grow constantly to reflect changes occurring both in language and in society. Over time, it will achieve a diachronic dimension to represent language used across generations. It will be used to find new words and phrases, locate newly coined terms, track variation in use of words and phrases, observe change in meaning, follow change in sentence structure, etc. The corpus developed by the DOE (Dash 2001) has the potential to be used as a general corpus provided it is allowed to grow at various dimensions with proper sampling and representativeness.

5.2 SPEECH CORPUS

Study on Indian speech has been confined within spoken English (Kachru 196, Bansal 1969, Nihalani at al. 1979). Nevertheless, we require speech corpora for all Indian languages. It will contain speech samples obtained from formal (e.g. radio and TV talks, interviews, news broadcasts, etc.) and informal talks (e.g. gossip, banter, etc.), impromptu conversations (e.g. quarrels, bargain, road-side talks, courtship babbles, hawkers' cry, political lectures, etc.), natural dialogues, and similar such speech sequences. It will be a reliable resource for both linguistic and non-linguistic information of a speech community, since it is capable to reflect on the core of a language revealing most of its characteristic properties. A speech corpus can record dialectal variations, phonemes and allophones, patterns of morphophonemic change in speech, and all demographic constraints that operate on speech.

Generation of comparable speech corpora in dialects and standard form will highlight differences existing between them concerning selection of phonemes, words, and sentences. Moreover, they will exhibit the nature/pattern of dialogic interactions noted in Indian speech projected in various linguistic discourses. In spite of such a rationale, we have not developed corpora in Indian speeches. Recently, the MIT has acquired a Hindi speech corpus that contains samples produced by a large number of speakers in variety of environments. It is now used as an essential data source for research and development activities in speech technology (Vikas et al. 2001: 33). Similar work is started for Bangla, Tamil, and Telugu. Other Indian languages are still in the queue.

In the present context of linguistic imperialism, generation of speech corpus in Indian languages will enable to preserve languages of the minorities, restore speech varieties,

and study speech patterns of the speakers coming from different walks of life. Linguistic analysis of such corpus will produce results that may be different form the results obtained from written corpus. Thus, speech corpus:

(i) Will reflect Indian languages as they are actually spoken in real life situations,
(ii) Will provide broad representative samples extending over wide selection variables (speaker's sex, age, class, etc.),
(iii) Will represent generalisations about spoken language as well as variations within spoken languages,
(iv) Will supply samples of naturalistic speech rather than speech developed under artificial conditions,
(v) Will furnish acoustic and phonetic aspects of speech research (important in telecommunication, voice-mail, etc.),
(vi) Will enable suitable quantitative analysis with phonetic and prosodic annotation,
(vii) Will encourage comparative study with written corpus to note primary similarities and differences.
(viii) Will provide relevant data and information for writing grammar of spoken language,
(ix) Will supply information for language teaching, speech processing, and other similar tasks.

Specialised speech corpora containing abnormal speech data is needed for analysing linguistic disorders of Indian people. It could be used to identify the patterns of linguistic impairment and the factors behind such disabilities. Findings could make valuable contribution in designing expert systems for repairing linguistic inability.

5.3 TAGGED CORPUS

We also need corpus tagged with extra-linguistic (e.g. text-type, publication year, name, age, sex of authors, domain, source and register of texts, etc.) and intra-linguistic (e.g. analytical marks, parts-of-speech codes, lexical category marks, grammatical category, sentence type, semantic and anaphoric codes, etc.) information. Though a general corpus has considerable value in language study and education, its utility is increased with annotation (McEnery and Wilson 1996: 24). Such corpus is useful for theoretical linguistic analysis (e.g. study of language and gender, language change, semantic change, ambiguity, etc.) as well as for designing language processing tools (e.g. word processor, morphological generator, POS tagger, sentence parser, spell-checker, lexical collocator, concordancer, lemmatiser, etc.). Recently, Electronic Research and Development Centre (ERDC), Noida, has tagged a Hindi corpus with parts-of-speech, which is integrated with translation system, language learning system, and spell checker and grammar checker software. Also, some written corpora are tagged to be used to design tools for word tagging, word count, letter count, frequency count, spell checkers, etc. (Vikas et al. 2001: 32). Similar effort is initiated to tag the MIT Bangla corpus at Indian Statistical Institute, Kolkata. We need to start encoding corpus of all Indian languages so that within a few years we can have tagged corpus for all linguistic activities.

5.4 OTHER CORPORA

Our present need in respect corpora in Indian languages is reported above. We believe these are the basic requirements that should be designed and generated with utmost importance and attention. All primary tasks of technology development can be initiated with these resources. However, after the generation of three types of corpora mentioned above, we can think of designing some other corpora (parallel corpus comparable corpus,

reference corpus, and special corpus) for their specific utilities and appellations. Among these, **parallel corpora** are valuable resource for comparative linguistic research and application between the source and the target language (Botley, McEnery, and Wilson 2000). We can develop parallel corpora among Indian languages (e.g. Hindi-Bangla, Bangla-Oriya, Telugu-Kannada, etc.) to extract information for machine translation, cross-lingual studies, inter-lingual communication (e.g. railway information, weather report, agricultural information, tourist information, government notice and circular, medical information, etc), cross-linguistic research, cross-language education, etc.

Generation of **comparable corpora** will allow us to find typological similarities among Indian languages as well as developing bilingual/multilingual grammars, lexicons, dictionaries, and other linguistic resources (Landau 2001: 273-342). Generation of language specific **reference corpora** will provide comprehensive information for writing grammars, making dictionaries, thesauruses, and language teaching materials. They will also supply extra-linguistic information to understand and evaluate various social and situational registers, demographic variations, cultural traits, and changes. However, the ultimate value of reference corpora lies in their ability in performing the role of a 'benchmark' for verification of LT tools and software. Formation of topic oriented, predefined, principle-based, and domain specific **special corpora** (e.g. corpus of dialects, child language, woman language, impaired language, slang, codes, non-native speakers, teenagers, games, auction, medicine, gambling, etc.) would contribute extensively in object-oriented and target-specific studies in India languages. The value of such corpora is appreciated in their ability in establishing uniqueness in representation of particular linguistic trait or feature within a specific domain of language use in society.

There is another issue related with the corpus generation in Indian languages. The present growth of corpus is now within the custody of the designers. Therefore, the resource is beyond the reach of the people who are not involved in corpus generation, but interested to use it in their work. However, in a recent development, both CIIL and MIT have put the data they have been able to obtain from various developing agencies, in the web for general access. We acknowledge their contribution. Yet, we advocate for the formation of a national archive for Indian language corpora, so that all corpora will be systematically preserved, documented, distributed, accessed, and utilised by the users. The Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), or International Computer Archive of Modern and Medieval English (ICAME) can be a good model for this. This central body will collect all corpora developed by individuals, institutes, and organisations; house them both in their archive; and arrange for making these available to the people or agencies involved with language related works.

6. CONCLUSION

The referential value of corpus is increased over the years due to technological growth and attitudinal change. It is manifested in new application of linguistics in telecommunication, information exchange, language technology, language education, and other fields. This will bring new opportunity to the Indian linguistics to survive. Utilisation of corpus in language study will yield new results to discard old linguistic theories, modify old practices of language study, and introduce empirical model of language research. What we need is a concerted effort for compilation of various representative corpora to use them as reliable databases. This is high time for us to turn our attention towards this. Else, we will be stumbling on the same old, non-reliable path of language study while others will run on the highway of corpus linguistics.

So far, we have argued for the initiation as well as continuation with the painstaking task of corpus generation in all Indian languages. If it is realised, by the end of this decade, we will be in a position to achieve both quantitative depth and qualitative width

in the generation of language resources, which are necessary to understand Indian life, language, culture, people, society, and heritage.

ACKNOWLEDGEMENT

Comments and views of W. Teubert of University of Birmingham, UK, P. Hall of Open University UK, B.B. Chaudhuri of ISI, Kolkata, and M.K. Nath of Calcutta University, Kolkata are acknowledged with thanks.

NOTES

[1] It may happen that there might be some more groups and organisations who are also involved in the same kind of works for Indian languages but are not referred to here. We admit that we could not mention their works due to the lack of information. We request them to inform the MIT, the governing authority of TDIL tasks in India about their present activities so that their contributions are properly acknowledged.

REFERENCES

Bansal, R. K. (1969) *The intelligibility of Indian English*. Monograph No. 4, CIEFL, Hyderabad, India.

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

Dash, N. S. (2001) *A Corpus-based Computational Analysis of the Bangla Language.* Doctoral Dissertation. University of Calcutta, Kolkata. (MS).

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.

Kachru, B. B. (1965) The Indianness in Indian English. *Word*. 2: 391-410.

Landau, S. I. (2001) *Dictionaries: The Art and Craft of Lexicography*. (2nd ed.) Cambridge: Cambridge University Press.

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

Murthy, B. K. and W. R. Despande (1998) Language technology in India: past, present, and the future. In the *Proceedings of the SAARC Conference on extending the use of Multilingual and Multimedia Information Technology (EMMIT'98)*. Pune, India.

Nihalani, P., R. K. Tongue, and P. Hosali (1979) *Indian and British English: A handbook of Usage and pronunciation*. New Delhi: Oxford University Press.

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.

Vikas, O., P. K. Chaturvedi, S. Lata, V. K. Sharma, and M. Jain (eds.) (2001) *Vishwabharat* (Indian Technology Newsletter 3), September 2001.

Vikas, O., P. K. Chaturvedi, V. K. Sharma, and M. Jain (eds.) (2002) *Vishwabharat* (Indian Technology Newsletter 4), January 2002.