Corpora in Indian Languages

B.D. Jayaram and K.S. Rajyashree
Central Institute of Indian Languages
Manasagangotri, Mysore-570 006, India.
Email: {jayaram, rajya}@ciil.stpmy.soft.net

1. Introduction

The status of Indian languages has changed from the time they became the official languages of the respective states and are used in wider domains like administration, mass media and higher education. The new status necessitated the development of different registers, writing of grammars and compilation of glossaries, dictionaries, encyclopaedias and thesaurus. Many governmental, semi and non- governmental organizations as well as individuals simultaneously got engaged in these activities. This led to diverse usage of lexical items and gave rise to competing forms in respective languages. The need was felt to standardize the usage, lexical items, grammatical constructions and discourse formats for given registers. The technological development to compile large corpus came in handy at this stage. The independence of the real examples from language is the strength of corpus. Apart from giving the real language usage, corpus can reveal collectively, aspects of language that are not obvious individually.

The official language status of Indian languages created need for language learning. Each state of India being plurilingual and increase in the interstate mobility the officials had to learn the official language of the state wherever they are posted. Also for mother tongue speakers of the respective languages, learning of language of different registers such as administrative or judicial language became necessary. Corpus of Indian languages could provide raw material for preparation of the need-based language learning materials. Also for those who learnt these languages in crash courses, computer based editorial aids like spell checker, grammar checkers or style checkers are of immense importance. The language corpus of running texts could be used to develop these editorial aids. The building of a corpus in major Indian languages was felt as need due to these factors and the Central Institute of Indian Languages took a lead in building corpora in Indian languages by coordinating with different institutions and by designing the nature and size of the corpora.

2. Corpus Design

The first issue considered for the nature of corpus is whether it should be spoken corpora or written corpora. The written corpora is favored because of practical difficulties in collection of spoken data as it involves more procedures, fieldwork, and availability of technical aids and trained linguists.

The issue about period of corpora had to be considered as modern Indian languages have a long history of about thousand years. The literature in most of the modern Indian languages dates back to 10^{th} century AD. While some of them even date back to 5^{th} century AD. The scope of the corpora is restricted to modern period, to be more specific, post independence period. The following factors are taken into consideration to arrive at this decision. First of all, after independence, the use of modern Indian languages was widened as they were used in domains like administration, judiciary, and higher education. Secondly, modern literaterary genera developed in most Indian languages around 1950 and various waves of modernity have been experienced by the different languages since then. After independence, when Indian languages assumed responsibilities as the official languages of the States, developmental programmes were taken to equip them with specialized vocabulary with either coining words, translating or borrowing. About three decades saw hectic activity by both governmental, semi and/or non-governmental agencies to develop different registers and to compile dictionaries, glossaries, thesaurus, encyclopaedias and grammars. A considerable amount of literature was produced in different registers during this period. After about 30 years modernity got consolidated in terms of vocabulary and standardized use of language. Taking into consideration all these points, it was decided to restrict

the corpora to the decade from 1981 to 1990, which would be representative contemporary Indian language corpora.

Once the period of corpora was decided the next point was to decide the type of corpora. There were two views: one was to develop a special purpose corpora either, generic, canonic, or period to start with and then gradually several such corpora together would develop into a large multipurpose corpora. While second view was to develop a multipurpose, general corpora at the beginning itself which will have a fair representation of all genera, authors, and periods of language. As the corpus was being developed for the first time and it was meant for any researcher working on modern Indian languages, it was felt appropriate to develop a general type corpora which would cater to multi-user. Further, for developing some of the tools like morphological analyzer, spell checkers, grammar checkers, electronic lexicons etc. data from different varieties of language are needed. However, a care was taken that if researcher is interested in working on a specific genera or register, the specific corpus should be available.

Before selecting the written texts, the point to be decided was whether it would include both printed and non- printed writings. The collection of non-printed manuscripts would involve certain type of problems and hence the corpus was restricted to printed written materials. Printed is further divided into categories informational, administrative, instructional, and imaginative. These four categories are subdivided further; e.g., informational is divided into learned, popular, and press news reports. Taking into consideration these points, the corpus is designed to have variety of genera and registers. 76 text categories are covered under six main categories namely, aesthetics, social sciences, natural, physical & professional sciences, commerce, official and media language and translated materials. (Appendix - I)

3. Corpus Size

The corpora are assumed to represent broad cross section of respective languages that would allow making generalizations about language. In this context, the relationship between the sample and the target population is very important. However, in case of language corpus the population to be sampled itself poses certain problems. As Clear (1992) puts it, in case of natural language, the phenomenon to be sampled is poorly defined. There is no obvious unit of language, which is to be sampled, which can be used to define the population and the sheer size of the population ensures that it will always be possible to demonstrate that some features of the population are not adequately represented.

Further, size of corpus may depend on the purpose of study. For example, in the beginning, the focus of corpus application was mainly on frequency and distribution of lexical items and hence both the American and British English corpora (Brown and LOB) which consisted of 500 samples of 2000 words each was considered large enough. In addition, since much attention was not given on the possible effect of the size of the corpus sample on the research results, the above sample size was considered sufficiently large. But, when the application of corpora data was extended to study syntactic structures, it was realized that even sample of 20,000 words was not sufficiently large to yield statistically reliable results (De Haan 1989). Haan subsequently pointed out that certain studies could be adequately undertaken on the basis of relatively small samples. However, there are certain studies, which are not only related to number of observations but also to the nature of the phenomenon studied and would require more sophisticated statistical techniques, need large samples for statistically reliable results (1992). Hence, there is no such thing as the best or optimum size of corpus.

Despite these difficulties, it is possible to have sampling design with theoretical basis taking into account pragmatic aspects. In the corpora of Indian languages the sample of written texts was obtained using stratified random sampling technique. The entire population of written texts is divided into maximum possible strata (text categories) and each stratum is given equal weightage for drawing sample. Taking into consideration the variables such as time, author, region/dialect and style (such as historical, social, mythological, etc.) wherever applicable the sample is drawn from each stratum. However, whenever the required data is not available under a particular stratum, that quantum of data is spread equally among all the major categories. This sampling scheme is considered appropriate because it is not possible to get equal representation from all the 76 categories, as the development of these categories in each language is not uniform. It can be seen from the discussion so far that we have not

strictly adhered to the theoretical principles of statistical sampling and inference, though some practical basis is established. As can be seen from the available natural language corpora, the principles of statistical sampling are not followed strictly in building a natural language corpus. (Jayaram, 1996)

However, there are many issues, which can be studied in details that would help in developing statistically more representative and reliable corpora. For example:

- The extent to which the selection of text within a genre actually represents the range of variation in that genre.
- The extent to which selection of genre actually represent the language
- Whether the text length adequately represents the overall characteristics of the text
- The overlapping of the text categories and genres

In the Central Institute of Indian Languages corpora are available in 12 Indian Languages, namely, Hindi, Marathi, Punjabi, Kashmiri, Urdu, Kannada, Tamil, Telugu, Malayalam, Oriya, Assamese and Bengali. Each language corpora has approximately three million words.

4. The purpose of Corpora

As has been mentioned earlier, the Indian language corpora built at CIIL are a multi-purpose corpora. The cost and time involved in compilation of a corpus makes it desirable to use the material as a basic resource for a wide range of varied linguistic studies. However, one must bear in mind that the corpus may not be the most appropriate sample for all the studies, which one might like to carry out. The more specific the purpose for a corpus, the better directed is be the data gathering.

The primary purpose of Indian language corpora is to provide following information:

- samples of words and sentences from different text categories
- frequency of words in different text categories
- frequency of grammatical categories
- occurrence of words in different context
- list of words in terms of grammatical categories
- collocation grammatical/syntactic patterning of individual lexical items
- raw material for the selection of natural examples of usage

5. Corpus Manager

A program is developed in Visual Basic to manage and maintain the corpora. The features of this programme are as follows:

- Viewing the data in respective scripts
- Retrieval of data in terms of categories and subcategories
- Retrieval of data in terms of fields like author and year
- Copying selected data to another destination
- Manual tagging
- Frequency of words in terms of file, sub-categories, major categories and entire corpus
- Word lists in terms of grammatical categories and its frequency
- Frequencies in terms of grammatical categories
- Length of sentence in terms of words
- Length of words in terms of syllables
- Concordance of words/ part of word with previous word, following word or complete sentence

The general type of corpora in Indian languages developed at CIIL is being used for the corpus-based language analysis that would be necessary for

- Compilation of different type of dictionaries
- Computer assisted language instructional material

- Domain specific machine translation
- Comparative linguistic study of different genres, registers, authors etc.
- Basic vocabulary of a specific language
- Disambiguation of meaning using statistical models (Jayaram, 1998)

The corpora in Indian languages is also being used for creating tools such as:

- Spell checkers, grammar checkers and style checkers
- Automatic editing
- Morphological parsers

The corpora in Indian languages are available free for any researcher for non-commercial use.

References

Clear, Jeremy, 1992. Corpus sampling in *New Directions in English Language Corpora* (ed.) Gerhard Leitner. Mouton de Gruyter, New York.

Haan, Peter de, 1992. The optimum corpus sample size? in *New Directions in English Language Corpora* (ed.) Gerhard Leitner. Mouton de Gruyter, New York.

Jayaram, B.D, 1996. *Development of Corpora in Indian Languages: Problems and Suggested Solutions*. Paper presented at Workshop on Indian Language Corpus and its applications at CIIL, Mysore.

Jayaram, B.D. and Umarani. P, 1998, Grammatical Category Disambiguation: A Probabilistic Model in *South Asian Language Review*, Vol. VII. No.1.

Appendix - I

Text Categories

Main Category	*Sub-category*
1**. Aesthetics**	
a. ***Literature***	1. Novel
	2. Short Story
	3. Essays
	4. Criticism
	5. Humour
	6. Children's Literature
	7. Biographies & Autobiographies
	8. Travelogues
	9. Letters/Diaries/Speeches
	10. Plays
	11. Science Fiction
	12. Folk Tales
	13. Text books (School)
b. ***Fine Arts***	1. Music
	2. Dance
	3. Drawing
	4. Sculpture
	5. Musical Instruments
	6. Hobbies
2. **Social Sciences**	1. Sociology
	2. Linguistics
	3. Psychology
	4. Anthropology
	5. History/Archeology/Epigraphy
	6. Political Science
	7. Home Science/Catering Technology
	8. Library Science
	9. Religion/Philosophy/Culture
	10. Economics
	11. Logic
	12. Journalism
	13. Folklore/Mythology
	14. Public Administration
	15. Law
	16. Business Management
	17. Education
	18. Text books (school)
	19. Demography

	20. Astrology 21. Criminology 22. Physical Education/Yoga 23. Health & Family Welfare 24. Forestry
3. **Natural, Physical and Professional Sciences**	1. Botany 2. Zoology 3. Geology 4. Geography 5. Bio-chemistry 6. Micro-biology 7. Physics 8. Chemistry 9. Mathematics 10. Statistics 11. Computer Science 12. Astronomy 13. Text books (School) 14. Medicine/Ayurveda/Homeopathy 15. Engineering 16. Architecture 17. Oceanography 18. Agriculture 19. Veterinary 20. Film Technology/Photography 21. Marine Biology/Fisheries
4. **Commerce**	1. Banking 2. Accountancy 3. Industry & Handicrafts 4. Finance
5. **Official and Media Languages**	1. Legal & Legislative 2. Administrative 3. Mass Media
6. **Translated Material**	1. Literature 2. Scientific 3. Legal 4. Administration 5. Education