CHAPTER II

Indian Elements in Current English: Methods and Materials

AIMS And OBJECTS

Many Indian words have entered into the English language during the last one hundred and fifty years. As stated earlier attempts have been made to identify the various words and specify the circumstances in which this has happened. However, there is a need for a comprehensive study of Indian words in the variety of English used in India i.e. Indian English and also see which of them have acquired currency in the native variety. The object of this study is to accomplish this to the extent possible.

The present study is based on 'The Kolhapur Corpus of Indian English' Shastri et al (1986). It may be appropriate here to discuss the concept of a corpus in some detail.

The Random House Dictionary of the English Language (1967) gives the following definition of 'corpus': 'A body of utterances or sentences assumed to be representative of and used for grammatical analysis of a given language or dialect'. W.N. Francis the chief compiler of the Brown Corpus-- broadens the definition to read: "A Collection of texts assumed to be representative of a given language, dialect or other subset of a language to be used for linguistic analysis". This way more facts are accounted for in a corpus such as "a corpus may be purposely skewed-- toward legal or scientific language-- and that it may be

used for phonological, graphemic, lexical or semantic as well as grammatical ahalysis" Francis (1979). The method of using a corpus was practised by lexicographers as early as in the 18th century and by writers of compendious grammars such as Jespersen, Visser, etc. Even for 'Grammar of Contemporary English' the monumental, authentic work produced in recent years by Quirk et al (1972) is based on the survey of English usage carried out at the University College London-- a corpus of written and spoken present day British English.

The first general purpose corpus of American English was compiled in 1961 at Brown University (Francis et al 1964). The compilers at the time hoped that it would serve as source material for all sorts of linguistic studies of American English -- lexical, grammatical, stylistic and so on. Within a decade of the building of the Brown Corpus, British scholars were attracted by the idea and a parallel corpus of British English the LOB corpus was built in the seventies at the University of Lancaster by Geoffrey Leech and others (Johansson 1978). The hopes of the compilers of these corpora may be said to have been more than fulfilled as we have over 500 scholarly studies on linguistic aspects of British and American English that have appeared (see ICAME News No.10 for a comprehensive bibliography).

Reviewing the practice of linguistic description Leech (1990) says that "there have been two highly influential and opposing views on the value of a corpus in linguistics over the past thirty or forty years. Firstly, post-Bloomfieldian structural linguists, such as Fries, Hill and Harris, regarded

the corpus as the only valid source of linguistic evidence, indeed as the fundamental reality which linguists had to describe. For them, intuition was an invalid source of evidence. Later, Chomsky and his co-workers turned this view upsidedown, by arguing that a corpus is of little or no value, and that the only sound source of evidence was the intuition of the native speaker. Since then, the Chomskyan view has persisted in practice, although it has been increasingly under attack from linguists".

Leech argues that "a corpus is important as a source--though not as the only source of evidence for linguistic descriptions". He suggests that "there is a kind of corpus evidence which is essential to linguistic competence of the native speaker, which is derivable from a corpus and which is <u>not</u> accessible to the unaided intuition of the native speaker". According to him, the importance of a corpus, as a basis for linguistic, study is self-evident.

All this he does in retrospect in support of his using the LOB Corpus for pointing out certain semantic nuances of the language exemplified in the use of certain pairs of synonyms such as 'almost' and 'nearly'.

Thus the use of corpus in linguistic description has gained ground once again. We have discussed the idea of Brown and LOB corpus of American and British English, as source material for linguistic study. Let us now turn to the Indian English Corpus. The first concerted effort towards a systematic and comprehensive description of Indian English may be said to be the

building of 'The Kolhapur Corpus of Indian English' parallel to the LOB and Brown Corpora of British and American English by Dr. S.V. Shastri in the early eighties. It is a million-word computer corpus of Indian English intended to be a representative, corpus of sample texts printed and published in 1978. The texts were largely selected by stratified random sampling process. The corpus consists of 500 texts of 2000 running words distributed over 15 genres of writing representing different styles. The composition of texts in the Indian Corpus is given in the table below:

Although the Indian Corpus is planned to be comparable to the Brown and LOB corpora there are some important differences dictated mainly by logistic and practical considerations.

The major departure is in respect of synchronicity. The Brown and LOB corpora draw their samples from the materials published in the year 1961, while the Indian corpus as stated earlier is drawn from materials published in the year 1978. It was felt that this decision would enhance the value of the Indian corpus as a source for the description of Indian English as the Indianness of Indian English is a post-Independence phenomenon. It is argued that in the same thirty years the American and British English may not have undergone such changes.

Table showing the basic composition of Indian English Corpus

	Text Categories	No. of texts in each category
A	Press : reportage	44
B	Press : editorial	27
C	Press : reviews	17
D	Religion	17
E	Skills, trades and hobbies	38
F	Popular lore	44
G	Belles lettres	70
H	Miscellaneous (Govt. documents, foundation reports, industry reports, college catalogue, industry house organ)	37
J	Learned and scientific writings	80
K	General fiction	58
L	Mystery and detective fiction	24
M	Science fiction	2
N	Adventure and Western fiction	15
P	Romance and love story	18
R	Humour	9
	TOTAL :	500

Materials and Methods

As mentioned earlier The Kolhapur Corpus of Indian English has been used as source material for the purpose of this study.

The corpus text follows a certain coding system characteristic of machine readable texts. Indian words, Indian expressions and hybrid expressions are coded as follows:

*4 prefixed to all Indian words

*5 ... *6 surrounds all Indian expressions

*(... *) surrounds all hybrid expressions

However, there is a certain amount of inconsistency in the coding of Indian English texts. While Indian words have largely been coded more consistently, the coding of hybrid expressions is rather very inconsistent. It appears from a cursory examination that hybrid formations with prefixes, and suffixes and hyphenated compounds have been adequately coded, hybrid formations consisting of a head and modifier have largely been left uncoded as such.

Given this corpus and its strengths and weaknesses we have made the best use of the material. To begin with all the Indian words marked *4, all the Indian expressions marked *5 ... *6 and all the hybrid expressions marked *(... *) were extracted by using the 'grep' utility on the UNIX operating system of the University Computer Centre. Then all the 'Indian words', 'Indian expressions' and 'hybrid expressions' were listed on cards.

i) An attempt was made to find out which of the Indian words have acquired currency in the native variety. For this purpose all the corpus words were checked against the entries in Webster's Third New International Dictionary of the English Language Unabridged (1961), (Web hereafter).

ii) Further an attempt was made to identify Indian words which do not occur in the corpus but which occur in Web.

As a result we arrived at three categories of words:

1) Indian words that occur only in the Corpus together with their frequencies (C category).

2) Indian words that occur both in the Web and in the corpus with their frequencies in the corpus (B category).

3) Indian words that occur only in the Webster's dictionary (D category).

Frequency figures were originally compiled from greped strings. Later the 'dictionary' of words in The Kolhapur Corpus compiled by Professor Gerhard Leitner was used to check and correct the figures.

iii) An attempt was made to arrive at some conclusions regarding non-occurrence of certain words in Web and non-occurrence of certain words in the corpus. One way of doing this was to group the words according

to semantic fields-- and see if there was any relationship between fields and occurrence/non-occurrence in Web/Corpus. So we classified these words into 27 categories, largely following Kachru (1975).

Category-wise -- i.e. (1), (2) and (3) above, and field-wise sorting was done by the Computer Centre.

iv) Further an attempt was made to find out semantic distribution i.e. whether the words in corpus are used in the same sense as that in Web. The details of our findings are reported in the following chapters.