PART II
CHAPTER IV

DATA AND METHOD OF ANALYSIS

4.1. Introduction

4.2. The Corpus

4.3. Limitations of the Corpus

4.4. Method of Analysis
CHAPTER IV

DATA AND METHOD OF ANALYSIS

4.1. Introduction

In Part II of this dissertation we present a description of the syntactic, lexico-semantic, idiomatic and collocational features characteristic of IE, based on the analysis of a corpus of written texts representing educated IE usage. The study also aims to enquire into the nature and possible sources of the deviations in IE with a view to acquiring a better understanding of the processes which have led to the emergence of this particular variety.

For our purposes, the term 'educated IE usage' refers broadly to the English of competent English-knowing Indian bilinguals, as opposed to varieties which are clearly substandard, such as Babu English, Butler English, Kitchen English, etc., as well as the 'interlanguage' varieties used by, say, undergraduate students still in the process of learning the language. We consider this variety among those most representative of IE because it is this variety that is ultimately acquired by the 'best' users of English in India. This variety may be identified in Kachru's terms as ranking somewhere between the central and ambilingual points on the cline of bilingualism.
Since the process of Indianization takes the form of a cline, we have further isolated three arbitrary points between the central and the ambilingual ones on this cline, each represented by a section of the corpus selected for our analysis. Our concern being to describe features that are pan-Indian, the data were drawn from a corpus of written texts representing as diverse an area within the country as was found reasonable for the limited purposes of this project.

As stated earlier (Chapter I, p.17), several studies of the whole range of varieties occurring at different points on the cline will be necessary before a complete description of IE is arrived at. This study is far from claiming any such descriptive exhaustiveness, but may be regarded as a modest step toward it.

4.2. The Corpus

The corpus selected for the present study contains samples of written texts of a non-fictional, semi-journalistic nature produced by educated users of IE. It is divided into three sections, corresponding to the three points isolated on the cline of bilingualism between the central and the ambilingual. The break up is as follows:
Section I: All the reading material from 6 issues of the Illustrated Weekly of India constituting a corpus of approximately 240,000 words.

Section II: 300 letters to the editor from newspapers reputed to have a national circulation constituting a corpus of approximately 60,000 words. The names of these newspapers and the number of letters taken from them are as follows:

- The Hindu (Madras) ... 150
- The Statesman (Delhi) ... 50
- The Times of India (Bombay) ... 50
- The Times of India (Aurangabad) ... 50

Section III: 300 letters to the editor from newspapers whose circulation is by and large limited to certain state capitals or large cities. The details of this corpus of 60,000 words are as follows:

- The Deccan Chronicle (Secunderabad) ... 75
- The Deccan Herald (Bangalore) ... 75
- The Poona Herald (Poona) ... 75
- The Searchlight (Patna) ... 75

In our total corpus, the material from the Illustrated Weekly is intended to represent nearly the highest point on the
cline of IE bilingualism. In choosing the Illustrated Weekly for this purpose we were prompted by the following considerations:

(a) Commanding a nation-wide circulation as it does, it draws contributions from different parts of the country, thus ensuring a truly representative sample of pan-Indian features.

(b) In view of the fact that it is one of the most prestigious magazines, those who contribute to it are likely to be among the most proficient users of IE.

(c) The nation-wide readership it caters to makes for a greater likelihood of the occurrence of pan-Indian features, as opposed to regional features. In order to be intelligible to as large a readership as possible, writers will tend to minimize the use of features that smack of a strong local or regional flavour - unless for a special effect.

(d) The fact that the Illustrated Weekly contains largely articles on such diverse subjects as politics, art, philosophy, science, history, education and culture and a variety of other topics of current interest, eliminates the danger inherent in analyzing data drawn entirely from IE fiction, where one is confronted with devices such as the deliberate creation of
'Indianisms' for special effect, conscious attempts to translate from the idiom of Indian languages, the writer's idiosyncratic style, etc. - all of which are perfectly legitimate in the context of creative writing in IE but not necessarily representative of the average educated Indian.

For these reasons the texts from the Illustrated Weekly form a larger chunk of the corpus (240,000 words). In addition to providing a fairly reliable and stable sample of data representing IE usage at this level, they also serve as a reference point with which to compare the samples in the two other sections. The latter consisting of letters to the editor from newspapers are intended to represent points on the IE cline lower than that represented by the material from the Illustrated Weekly. The material from the national dailies may be said to represent a slightly higher point on the IE cline than that represented by the material from regional newspapers.

This corpus from the newspapers is comparatively smaller in size - 60,000 words from the national dailies and another 60,000 words from the regional dailies. The primary objective of including this material in our corpus is to examine what some of the features of IE are whose incidence increases as we
move down the cline of IE. This may help us in pinpointing some of the characteristic tendencies of the written variety of IE.

There is also another reason for including these letters to the editor in our corpus: these letters constitute genuine samples of the written English of educated Indians, who are not particularly trained to write professionally. These letter-writers write to the editor on matters of interest to them. The fact that they choose to write in English shows that they feel sufficiently at home with the language to voice their views in it on matters about which they feel most strongly. While in section II the occurrence of regional features would continue to be— at a minimum, since the readership at this point is still a 'supra-regional' one, in section III we expected the occurrence of regional features to be greater. The fact that the data show a general increase in the occurrence of certain kinds of Indianisms seems to corroborate this assumption.

The following considerations made us choose a corpus of written materials for this project, though we do believe that a complete description of IE must deal with the spoken as well as the written forms of this variety:

(i) Writing demands and gets more careful attention. Whereas in speech one thinks as one goes along, in writing one
has a lot more time to think, to rephrase one's sentences before committing them to paper. In this sense a written corpus should provide a closer and more reliable indication of the competence of IE users by eliminating the occurrence of deviations in speech which are merely performance errors.

(ii) For most Indians contact with the English language has been through the written mode rather than the spoken one. Besides, since English in India is used more frequently in writing than in speech, many Indians feel more at ease in this medium and the formal styles that go with it. It therefore appears more reasonable to choose the written mode as a basis for description.

4.3. Limitations of the corpus

A limitation of the kind of corpus selected here is that such material has normally to undergo editorial scrutiny before it is published. Since there is a possibility of many 'Indianisms' being eliminated in the process, the question may be asked as to how truly representative of IE such a corpus is likely to be after the editor has run his pencil through it. However, the seriousness of this objection is minimised to a great extent if we consider the fact that those responsible for editorial scrutiny are not native speakers of British or American English, but themselves speakers of IE.
In fact, the above mentioned limitation could be turned to advantage if we accept that the Indianisms that persist in the corpus even after it has passed muster have done so precisely because they failed to strike the editor as odd or peculiar in any way, and seemed perfectly acceptable in the Indian socio-cultural context. That is to say, what survives, survives either because it is unobtrusive or because it is necessary or because it is not entirely idiosyncratic. The retention of these Indianisms therefore, may be an indication of their being among the more stable and permanent features of IE - and hence a truer reflection of the competence of IE users.

Further, our intention in selecting a limited body of corpus was to obtain a fairly general understanding of the structure of IE at certain points on the IE cline; a complete coverage of IE usage will require a large scale survey. In other words, the features described as 'Indianisms' are characteristic, but not wholly constitutive, of IE. It is conceivable that examination of additional chunks of data may yeild an extra set of features in addition to those isolated in this study, though this seems unlikely, because as we looked at more of the same kind of data, the configuration of features distinctly Indian seemed to remain the same, though
individual occurrences of these features increased. In any case, it must be noted that, in a sense, the texts examined here are to be regarded as only approximatively representative of the IE speech community, and that the adequacy and reliability of the description will be proportionate to the adequacy and reliability of the data examined here.

4.4. **Method of Analysis**

In this study we have treated IE as a non-native dialectal variety of English. Dialect studies involve dealing with groups of people that are in some way different from each other. The difference may be pre-dominantly geographical, or it can be socioeconomic or even ethnic. Each point of view is an abstraction based on a collection of differing speech patterns that share "a nameable commonality" (Wolfram 1974:9). Non-native varieties such as IE present an additional dimension of complexity in that they are resultant of processes of second language learning as well as languages in contact.

On the analogy of dialect studies, two alternative methods of describing non-native varieties suggest themselves. One method is to write a separate grammatical description of every variety/dialect, starting from scratch, as linguists do when
confronted with new and unknown languages. However, as is evident, this approach is highly uneconomical for our purposes here. The writing of separate grammars is a case which may be justifiable for substantially divergent varieties such as pidgins and creoles. But writing a separate grammar for the subvarieties of a single code seems a wasteful and unnatural procedure since they all tend to share a substantial part of their syntax. This is true at least of educated written Indian English vis-a-vis any native standard variety.

The other alternative is to assume a 'common core' or nucleus that all varieties of the language share, i.e., as Quirk et al. (1972:14) put it, to assume that however esoteric or remote a variety may be, it has running through it a set of grammatical and other characteristics that are present in all others. In addition to the 'common core', of course, every variety has features which are peculiar to it and which serve as formal (and sometimes substantial) criteria or markers of the variety in question.

We have treated IE in a similar fashion - assuming a 'common core' which it shares with all other Englishes - native and non-native. Features belonging to the common core are unmarked. Those that do not belong to the common core, but have
become a fairly stable and institutionalised part of IE are deviant or marked and treated as characteristic of IE.

The Indianisms in our data were for the most part identified with the help of a native speaker of (Standard British) English* who meticulously examined the major bulk of the corpus and corroborated our identification of Indianisms by pointing out those features which struck him as syntactically, lexically, idiomatically and collocationally deviant - i.e. as "unEnglish" English and thereby as markedly Indian. It was hoped that the basis of our study would be made more valid if a native speaker's judgements as to what constitutes IE were used to confirm our own, though the question remains as to whether the intuitive judgements of a single native speaker are sufficiently reliable for this purpose. Since it was not possible, under the circumstances, to get a larger number of native speakers to examine such a vast corpus, however desirable, other native speakers of English were consulted only on points of uncertainty. Although relying on the intuitions of a British English speaker implies, in a sense, taking this variety as a point of reference, the fast-spreading

---

* Much of the validity of the analysis in this study derives from consultations with Mr. R.K. Tongue, British Council expert and author of *The English of Singapore and Malaysia* (1974), who worked at CIEFL, Hyderabad, during 1974-77, and who gave valuable guidance in isolating the 'deviations' in the corpus and discussed at length individual examples and also their possible root causes.
influence of American English on IE has also been taken into account. And the fact that this particular native speaker has lived and worked outside of England as a British Council ELT  
expert ensured to some extent, his familiarity with other Englishes as well.

In classifying the deviations each point of grammar, usage or style violated was considered a separate instance of deviation. For example, if an utterance like

The traffic police should lose no time to book cases against vehicles emitting black smoke which not only causes air pollution and are a serious health hazard besides being contributive in one way or the other for road accidents

was found to diverge from native English in, say, 5 different areas of syntax (as has been indicated by the underlining above), we considered it an instance of 5 deviations and not one. These deviations were classified as follows:

(i) the 'to' infinitive instead of preposition + gerund
to book cases → in booking cases

(ii) violation of subject-verb concord
black smoke ... are → black smoke ... is

(iii) tendency towards N + N constructions
health hazard → hazard to health
(iv) **lack of cohesion due to clumsy syntax**

More specifically, faulty use of intra-sentential conjunctions. In "... not only causes air pollution and are a serious health hazard besides being contributive ..." the coordinating correlative 'but also' has been omitted and 'besides' has been used instead. A less serious divergence is the use of non-parallel conjuncts with 'and', where a lexical verb (cause) occurs in one, and in the other, a form of the verb 'be'.

(v) **substitution of preposition**

contributive ... for → contributive ... to

Our method of processing the data may superficially look like error analysis. We have already explained earlier (Chapter I, p.27) how our procedure and intention are entirely different from those of error analysis.

After a careful scrutiny of the data those linguistic items which were identified as marked in this manner and therefore characteristic of IE, were classed under three major areas: I syntax, II lexis, collocation and style/register, and III cohesion. These were then classified into 21 main categories in terms of the differences observed in grammar and usage. Some of these categories were further subclassified.
Given below is a list of the main categories and the sub-categories used in our analysis:

I. Syntax

1. Articles
   a. Redundancy in the use of the definite article
   b. Omission of the definite article
   c. Redundancy in the use of the indefinite article
   d. Omission of the indefinite article
   e. Use of the definite article for the indefinite article
   f. Use of the indefinite article for the definite article

2. Prepositions
   a. Substitution of certain prepositions
   b. Redundancy in the use of prepositions
   c. Omission of prepositions
   d. The use of 'due to' as a compound preposition

3. Adverbial Particles
   a. Substitution of certain adverbial particles
   b. Redundancy in the use of adverbial particles
   c. Omission of adverbial particles
(iv) lack of cohesion due to clumsy syntax
More specifically, faulty use of intra-sentential conjunctions. In "... not only causes air pollution and are a serious health hazard besides being contributive ..." the coordinating correlative 'but also' has been omitted and 'besides' has been used instead. A less serious divergence is the use of non-parallel conjuncts with 'and', where a lexical verb (cause) occurs in one, and in the other, a form of the verb 'be'.

(v) substitution of preposition
contributive ... for contributive ... to

Our method of processing the data may superficially look like error analysis. We have already explained earlier (Chapter I, p.27) how our procedure and intention are entirely different from those of error analysis.

After a careful scrutiny of the data those linguistic items which were identified as marked in this manner and therefore characteristic of IE, were classed under three major areas: I syntax, II lexis, collocation and style/register, and III cohesion. These were then classified into 21 main categories in terms of the differences observed in grammar and usage. Some of these categories were further subclassified.
Given below is a list of the main categories and the sub-categories used in our analysis:

I. Syntax

1. Articles
   a. Redundancy in the use of the definite article
   b. Omission of the definite article
   c. Redundancy in the use of the indefinite article
   d. Omission of the indefinite article
   e. Use of the definite article for the indefinite article
   f. Use of the indefinite article for the definite article

2. Prepositions
   a. Substitution of certain prepositions
   b. Redundancy in the use of prepositions
   c. Omission of prepositions
   d. The use of 'due to' as a compound preposition

3. Adverbial Particles
   a. Substitution of certain adverbial particles
   b. Redundancy in the use of adverbial particles
   c. Omission of adverbial particles
4. **Tense**
   a. Tendency to use the past perfect tense in place of the simple past
   b. Tendency to use the present progressive tense in place of the simple present
   c. Incorrect selection of tense
   d. Incorrect sequence of tenses

5. **Modals**

6. **Word order**

7. **Non-count nouns used as count nouns**

8. **Plural form with certain words**

9. **Confusion between gerund and 'to' infinitive**

10. **Verbal '-ing'**

11. **'That' clause after certain verbs**

12. **Transitive/intransitive verbs**

13. **Reflexive/non-reflexive verbs**

14. **Modification**

15. **Intrusive 'as'**

16. **N + N type constructions**

II. **Lexis, Collocation and Style/Register**

17. **Lexis**
   a. Unidiomatic expressions
   b. Substitution of lexical items
   c. The use of synonyms and quasi-synonyms
d. Pompous, high-sounding, bookish words

e. Malapropisms

f. Violation of selectional restrictions

g. Extension of derivational processes

h. Neologisms

i. Semantic extension

j. Semantic restriction

k. Semantic transfer

l. L₁ translations

m. Typical SAE usage

n. Borrowings

18. Collocation

a. Extension of the range of collocability

b. Notionally related words

c. Pompous, high-sounding, bookish words

d. Blending

e. Mixed metaphors

f. L₁ translations

19. Set Phrases and Idiomatic Expressions

20. Style/Register

*SAE = South Asian English.
III. Cohesion

21. Absence of Cohesion
   a. Absence of specific anaphoric reference, ambiguous and partial reference
   b. Inappropriate sentence connectors
   c. Non-parallel conjoined clauses/phrases
   d. Inappropriate selection of conjunctions
   e. Excessive co-ordination
   f. Inappropriate selection of complementizers
   g. Excessive subordination
   h. Dangling participles
   i. Fusion of two alternative phrases
   j. Ellipsis
   k. Pleonasms and superfluous constructions
   l. Condensation
   m. Awkward expression resulting from clumsy syntax
   n. Awkward expression resulting from L1 translations
   o. Longwindedness

The results of this analysis are presented in the following chapter. For each subcategory illustrative examples drawn from the corpus are given.

After classifying the marked linguistic items into the categories and sub-categories mentioned above, we made a count
of items in each of these categories and subcategories. In all 3236 deviations were recorded in the corpus. Deviations classified under lexis and collocation, and lack of cohesion account for more than 50% of the total deviations recorded in the whole corpus. A statistical summary of the type and number of deviations is given in Chapter VI.