Chapter Three

Research Design and Methodology

Contents

3.1. Introduction
3.2. Research objectives
3.3. Subject selection
3.4. Writing task
   3.4.1. Administering the writing task
   3.4.2. Rating the essays
   3.4.3. Entering the study texts into the computer
3.5. Instrumentation
   3.5.1. Coh-Metrix
      3.5.1.1. General identification and reference information
      3.5.1.2. Referential indices
      3.5.1.3. Coreferential indices
      3.5.1.4. Connectives
      3.5.1.5. Conceptual coherence
      3.5.1.6. Situation model indices
      3.5.1.7. Other measures
      3.5.1.8. Procedure
   3.5.2. Topical Structure Analysis
3.6. Data Analysis
CHAPTER 3
RESEARCH DESIGN AND METHODOLOGY

3.1. Introduction

This study was an attempt to investigate the features of local and global coherence in the expository essays written by Iranian EFL and Indian ESL undergraduate students in terms of the degree to which these two groups of students achieve coherence in their compositions and the types of coherence devices they use in their essays. Moreover, the researcher tried to study the ways in which these ESL and EFL writers develop their topics throughout their paragraphs to achieve internal coherence. To this end, the expository essays written in English by two groups of undergraduate students from Iran and India were compared in an effort to attain the following research objectives and test their relevant null hypotheses:

3.2. Research objectives

This study aimed at achieving the following four research objectives:

1) The study aimed at investigating any possible differences between the Iranian EFL and the Indian ESL undergraduate students with regard to their use of explicit referential and conjunctive cohesive features as measured by the computer application Coh-Metrix. These features include:
   a) Anaphoric references (at adjacent and distant sentence levels)
   b) Sentence connectives (incidence of all connectives in general and the different types of connectives in particular)

2) The second objective of the study was to find out if there were any differences between the Iranian EFL and Indian ESL undergraduate students regarding their use of lexical and coreferential ties as measured by the computer application Coh-Metrix. The lexical devices used include:
   a) Argument overlap (at adjacent and distant sentence levels)
   b) Word-stem overlap (at adjacent and distant levels)
   c) Content word overlap
d) Latent semantic analysis or semantic similarity (at adjacent and distant sentence levels)
c) Mean hypernym value of nouns

3) The third objective set for this study was to investigate any differences in the types of situation model dimensions created in the English texts written by the Iranian EFL and Indian ESL undergraduate students as measured by the computer application, Coh-Metrix. These dimensions are as follows:
   a) Causal dimension (created by causal verbs, links and particles)
   b) Intentional dimension (created by goal-directed actions with animate agents)
   c) Temporal dimension (created by sequence of verbs classified in a particular tense and aspect)
   d) Spatial dimension (created by location nouns and prepositions as well as motion actions and prepositions)

4) The fourth objective of the study was to investigate possible differences between the Iranian EFL and Indian ESL undergraduate students with regard to the types of topical progressions they employed in their paragraphs. Three types of progressions were targeted:
   a) Parallel progression
   b) Sequential progression
   c) Extended parallel progression

The design of this research work consists of two major components:
1) a web-based text analysis tool called Coh-Metrix and
2) Lautamatti’s (1978) Topical Structure Analysis (TSA)

Coh-Metrix is a web-based computer application that can be used in many different ways to investigate the cohesion of the explicit text and the coherence of the mental representation of the text. This software is capable of analyzing texts for over 200 measures of cohesion, language, and readability. Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics. After the user enters an English text, Coh-Metrix returns measures requested by the user (Graesser, et. al, 2004).
Lautamatti’s (1978) Topical Structure Analysis (TSA) involves the manual analysis of texts to see how topics progress at sentence level and across sentence boundaries. These instruments will be discussed in more details later.

The present chapter consists of four main sections. The first section deals with the process of subject selection, the second section addresses the writing task and procedure, the third section describes the procedures involved in the analysis of the texts, and the last section discusses data analysis method outlining the statistical techniques used in order to test the research hypotheses.

3.3. Subject selection

Subject selection for this study followed the ‘area or cluster sampling’. This method of sampling is more convenient and more economical to sample a large population. According to Best (2003: 31), “The area or cluster sample is a variation of the simple random sample that is particularly appropriate when the population of interest is infinite, when a list of the members of the population does not exist, or when the geographic distribution of the individuals is widely scattered.” The Iranian subjects were randomly selected from among the students studying in two Iranian universities: 1) Roudehen Azad University in the capital, Tehran and 2) Persian Gulf University (PGU) in Boushehr. However, since it was suspected that in India the researcher might not have been able to have access to enough subjects with the same linguistic background, he decided to select more colleges for sampling. Therefore, four colleges affiliated to Panjab University in Chandigarh were chosen: 1) Government College for boys, 2) Sri Guru Gobind Singh College, 3) MCM DAV College, and 4) DAV College for boys.

The primary sample of this study consisted of 295 subjects from Iran and India. A total of 172 Indian and 123 Iranian senior undergraduate students writing in English participated in the present research study. In order to minimize inter-subject variations, the subjects were selected based on their level of education; therefore, the sample was restricted to senior students on the assumption that these students, whether Iranian or Indian, have acquired the necessary writing skills in English.

The Iranian sample comprised 43 male and 80 female students, and the Indian sample consisted of 56 male and 116 female students. The following table presents the
primary statistics regarding all the subjects participating in the study:

**Table 3.1: Statistics for Iranian and Indian subjects**

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian</td>
<td>56</td>
<td>116</td>
<td>172</td>
</tr>
<tr>
<td>Iranian</td>
<td>43</td>
<td>80</td>
<td>123</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>196</td>
<td>295</td>
</tr>
</tbody>
</table>

Unlike Iranian students who spoke Persian as their mother tongue, the Indian subjects came from different linguistic backgrounds, such as Hindi, Panjabi, Manipuri, etc. Among the Indian subjects 92 students came from Hindi background, 57 spoke Panjabi as their vernacular, and 23 came from other language backgrounds. The following table shows the statistics for the Indian subjects:

**Table 3.2: Statistics for Indian subjects**

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>37</td>
<td>55</td>
<td>92</td>
</tr>
<tr>
<td>Panjabi</td>
<td>10</td>
<td>47</td>
<td>57</td>
</tr>
<tr>
<td>Others</td>
<td>9</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>116</td>
<td>172</td>
</tr>
</tbody>
</table>

In order to make comparing of the essays possible, they had to be grouped on the basis of the subjects’ nationality, vernacular and gender. This was very crucial to the analysis process because only those subjects who shared the same mother tongue were to be compared with their counterparts in the other group. Moreover, the researcher wanted to ascertain that there were a proportionate number of boys and girls in each group. After careful examination of the samples, it was found out that among the Indian subjects only those who spoke Hindi as their mother tongue lived up to the research criteria and could be compared with the Iranian students. Thus, the rest of the subjects, including Panjabi and other students were excluded from this study. There were also 17 essays in the Iranian group and 19 in the Indian group which were regarded as off-topic, so they were left out.
This stage of the research left a total of 179 essays written by the Indian and the Iranian students for later analyses. The Iranian subjects included 34 male and 72 female students speaking Persian as their mother tongue, and the Indian subjects comprised 30 male and 43 female students who spoke Hindi as their native language.

Two points regarding the samples should be noted here. First, the number of girls and boys in each group is representative of the original populations. That is to say that originally the girls in both sample populations outnumbered the boys. Second, the unequal size of the samples does not have any adverse impact on the statistical results because the number of the subjects in each sample exceeds 30.

Table 3.3: Summary of the statistics for the Indian and Iranian subjects:

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian</td>
<td>30</td>
<td>43</td>
<td>73</td>
</tr>
<tr>
<td>Iranian</td>
<td>34</td>
<td>72</td>
<td>106</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>115</td>
<td>179</td>
</tr>
</tbody>
</table>

3.4. Writing task

The first step in the writing process was to assign a task that would meet the three requirements of context of situation and register put forth by Halliday and Hasan (1976). They maintain that for a string of sentences to be considered as text and make sense, it should meet the requirements of context of situation which is characterized by the following features:

- **FIELD**: the total event, in which the text is functioning, together with the purposive activity of the speaker or writer; it thus includes the subject-matter as one element in it.
- **MODE**: the function of the text in the event, including therefore both the channel taken by the language—spoken or written, extempore or prepared—and its genre, or rhetorical mode, as narrative, didactic, persuasive, phatic communion, and so on.
- **TENOR**: the set of relevant social relations, permanent and temporary, among the participants involved (p. 22).
Thus in order for the essays to meet the first requirement of register, the FIELD, the subjects were asked to write an essay describing an event and its meaning, to meet the second requirement, the MODE, the essays were to be written in the form of exposition using any technique so they can convey any necessary information about the event and its meaning to their audience. And finally, for the essays to meet the last requirement, the TENOR, the students had to address their fellow students and their teachers. As it is very likely that the topic of an essay affect its structure, the subjects were asked to write on a predetermined familiar topic:

Describe three ways in which the internet affects your life as a student.

This topic was decided on the assumption that it was familiar and easy enough for the subjects to write about as much as needed. Second, it was expository in nature; it imparted some information to the readers by describing rather than arguing to prove or refute a claim.

3.4.1. Administering the writing task

The subjects from Iran and India were given the same topic mentioned above to write about. There was one class-period time limit imposed, but there was no constraint on the number of words so that they can write as many words as possible to express themselves and develop the topic well enough. They wrote the essays in the classroom under the supervision of their teachers. They were also asked to mention their names, gender, and mother tongues.

3.4.2. Rating the essays

The first step in the analysis was rating the essays. The essays were evaluated based on the CUNY six-point holistic scale for their ability to address the topic and take a position, organization and development of ideas, and competency in Basic English sentence structure, word usage, and mechanics. Each score reflects an evaluator's overall judgment of the writer's performance in relation to the skills identified above. The writings were rated as exceptional (6), superior (5), competent (4), adequate (3), weak (2), and inadequate (1).

After the essays were scored, they had to be re-evaluated for inter-rater reliability. To this end, 62 (or 30%) of these essays were randomly picked out for re-evaluation.

86
Besides the researcher, two other raters from Iran and India were involved in scoring the compositions. They had years of experience in teaching English to Persian and Indian students. Prior to the re-evaluation process, the raters were instructed on how to score the essays based on the CUNY six-point-scale, and a copy of the instructions were also attached to the papers for quick and easy reference. The result of the correlation test indicated a coefficient of .80 which was significant at P < .01 denoting a high inter-rater reliability estimate.

3.4.3. Entering the study texts into the computer

Entering the texts into the computer had two considerable advantages to the present study. First, the essays had to be analyzed by a computer software called Coh-Metrix, so in order for the software to be able to read the essays and analyze them, they had to be typed. Secondly, the study also involved a manual analysis of the essays using Lautamatti’s (1978) Topical Structure Analysis (TSA) model. This required coding and recoding of the texts for several variables, and this could lead to illegibility and obscurity of the handwritten texts; therefore, typing the essays eliminated this problem and enabled the researcher to produce as many copies of the essays as needed for the pertinent coding.

The process of entering the texts was an onerous and time-consuming task. Special care had to be taken for the essays to be typed in the same way as they appeared in the handwritten form, i.e. no grammatical or spelling error had to be corrected. Reading the handwritten copies was also a rather challenging endeavor. Therefore, to ensure the accuracy of the typing process, the researcher sought the cooperation of a professional typist. During the typing period the auto-correction and grammar and spell check of the Microsoft Word remained deactivated so that it would not automatically correct errors while typing. After all the essays were typed, they were closely rechecked for any possible mismatch between the handwritten and typed versions of the essays.

At this point, the essays were almost ready for analysis. Two instruments were used for the analysis of the essays: 1) Coh-Metrix software, an automated tool and 2) Lautamatti’s (1978) Topical Structure Analysis (TSA) model, a manual tool.
3.5. Instrumentation

The developments in the fields of corpus linguistics and computational linguistics have led to the creation of computer resources and applications that could help analyze texts with reasonable speed and accuracy. For instance, huge corpora of written and spoken language have been built and made accessible to researchers. Concordancers were developed so that they can create an index of all the words in a text or a body of texts. Syntactic parsers and part of speech classifiers have been made so that they can annotate and classify parts of speech in text. Statistical representation of world knowledge (Latent Semantic Analysis) has been made possible in order to show how conceptually and semantically similar and coherent different segments of a text are. Furthermore, the importance of coherence and cohesion of text and their impact on text quality, readability and comprehensibility of text have led some researchers in different fields of linguistics to focus on these features and the ways they can be analyzed. They tried to think up tools and models that can explain and analyze them. One of these tools is a web-based software called Coh-Metrix (which was used for the analysis of the students’ essays in this study).

3.5.1. Coh-Metrix

Coh-Metrix is a web-based software tool which is capable of analyzing and assessing text quality and cohesion on over 50 types of cohesion relations and more than 200 measures of language, text, and readability. It was originally meant to replace readability formulae to assess text difficulty, but now it has proved its applicability in some other textual analyses (e.g., Dufty et al 2005, McCarthy 2006b, Hall, et al. 2007; McCarthy, 2006; McCarthy, 2007). Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, statistical representations of world knowledge (Latent Semantic Analysis), and other components that are extensively used in computational linguistics (Graesser et al., 2004; Louwerse et al, 2004).

Coh-Metrix application operates at two central levels of discourse processing: 1) text-based cohesion and 2) mental-model based coherence. According to recent theories of discourse, a text has at least three levels of structure: the surface code, the textbase, and the situation model. The surface code refers to the literal words which are used in the text, the textbase includes the propositions that the surface code describes, and the situation
model consists of the representation of the world that the text is intended to convey. Therefore, Coh-Metrix is developed so that it can capture many of the features of textual cohesion and its aspects, coherence, at all the three cognitive representation levels that get activated during reading comprehension: surface code, textbase, and the situation model (McNamara, Louwerse, and Graeser, 2002).

A demo version of this software as well as the detailed information about it is available for researchers on its website at (www.cohmetrix.memphis.edu). It provides almost 50 researched indices of textual cohesion and coherence. The results offered by the demo version are as valid as those provided by the full version (McCarthy, personal communication, 2008). The primary information provided by the Coh-Metrix is related to General identification and reference information.

3.5.1.1. General identification and reference information

Once the software is fed with a given text through a copy-paste facility, it will return all the identifying information provided by the user at the time of entering the text like the title he gives to the text, the source from which he has taken, the JobCode he assigns to it for later retrieval or viewing the output data, the genre or the general category of the text he has selected for later reference such as science, narrative, or information, and the LSA (semantic or the world knowledge space) he selects such as college level, narrative, information etc. Then the main results of the analysis follow.

The results of the text analysis by this software are given in indices. Some of the indices measure the same thing; therefore, in this study the researcher used only those indices that best suit the aim of this research work. They not only capture much of Halliday and Hasan’s taxonomy of cohesion but also include interesting features that are indicative of more semantic and global coherence of the text. The indices used in this work are as follows: Referential indices, Coreferential indices, Connectives, Latent semantic analysis (Indices of semantic or conceptual coherence indices), and Mental or situational model dimensions. These indices will be discussed below.

3.5.1.2. Referential indices

A class of indices which are measured by Coh-Metrix and play a crucial role in cohesion and coherence of a text involves referential and semantic indices. Referential
cohesion occurs when a noun, pronoun, or a noun phrase refers to another constituent which is regarded as the referent of the noun. Depending upon whether the referents proceed or follow a given referring item, they are divided into anaphoric or cataphoric referents (Halliday and Hasan, 1976). Anaphoric reference occurs when the referent appears at an earlier point in the text, for example,

My friend immigrated to Canada a few years ago. He has started a big business there.

Cataphoric reference, however, takes place when a referring item, like a pronoun, looks ahead at its referent in the text, for instance,

After he immigrated to Canada a few years ago, my friend started a big business there.

This software measures the degree to which the sentences in a text are anaphorically related both in terms of adjacent and distant sentences.

The incidence of pronouns and their ratio in a text can be very important; they can add to and facilitate textual cohesion of the text, if they are used reasonably; otherwise, they can lead to more confusion and ambiguity if the reader cannot understand what the pronouns refer to.

3.5.1.3. Coreferential indices

Coh-Metrix also measures local and global coherence of a text in terms of three forms of coreference: argument overlap, noun overlap and stem overlap. Argument overlap occurs when a noun, pronoun, or noun phrase in one sentence is the coreferent of a noun, pronoun, or noun phrase in another sentence. For instance, in the following example all the underlined words are coreferent.

My grandparents live in New York. They often come and see us, but this year, due to my grandma's illness, they could not pay us a visit.

There is noun overlap if two sentences share a common noun with the same morphological stem.

There is much discussion in the academic world of communication as to what actually constitutes communication. Currently, many definitions of communication are used in order to conceptualize the processes by which people navigate and assign meaning. Communication is also understood as the
exchanging of understanding.

In fact, noun overlap is the result of the overlap of two nouns with the same morphological stem. However, if there is a noun in one sentence which has a stem shared by any word category (such as a noun or a verb) in another sentence, then the two sentences have stem overlap (Graeser et al, 2004), for example,

They were first introduced by a friend. They never imagined that this introduction would lead to a lifelong friendship.

In the above sentences, the words “introduce” and “introduction” share the same word stem, hence, they lead to more cohesiveness of the text.

3.5.1.4. Connectives

Connectives are words that linguistically signal coherence relations between text segments. They play a facilitative role in reading, save reading time, and improve comprehension (Degand, Lefevre and Bestgen, 1999). Nonetheless, careless and overuse of these linguistic devices can lead to the false illusion of understanding the text and interfere with the readers’ inference making and generation of elaborations beyond explicit connectives. In other words, overuse of these connectives in the text may increase readers’ passivity. The readers prefer to have recourse to these cohesive signals than using their background knowledge and inference making ability to establish coherence between text segments (Noordman et al. as cited in Degand et al. 1999).

Coh-Metrix software can identify, count and classify the connectives in the text. First, it will give the total number of connectives used, and then it will classify them based on two dimensions: on one dimension they are divided based on whether they extend the situation described in the text or not. In this respect they are categorized into two groups: 1) positive connectives that extend the situation and 2) negative (synonymous with adversative) connectives which cease to extend the situation. For instance, words like and, after, and because are regarded as positive and but, until and although are regarded as negative (adversative) connectives. On the other dimension, connectives are grouped based on their types: 1) additive, 2) causal, 3) logical, and 4) temporal. Examples of these types of connectives are as follows:

Additives: also, moreover, and, etc.

Causal: because, so, consequently, although, nevertheless, etc.
Logical: or, actually, if, etc.
Temporal: after, before, when, until, etc.

This computer application also identifies and measures the logical operators and sentence syntax similarity. Logical operators are words like and, or, not, if, and then, and their variants which usually occur in texts that express logical reasoning. They can be applied to phrases (verb phrases and noun phrases), clauses, and sentences. The presence and the density of these operators may result in more analytic texts and create difficulty in comprehension for some readers.

3.5.1.5. Conceptual coherence

Another method of computing semantic and conceptual similarity is Latent Semantic Analysis. Latent Semantic Analysis (LSA) is a statistical technique for representing world knowledge based on a large corpus of texts. Unlike other referential overlap indices which focus on pairs of words that share common stems like communicate/communication, LSA extends referential overlap beyond such explicit relations like stem overlap or noun overlap into relative semantic similarity between words that enjoy shared conceptual similarity such as computer/internet, internet/network, network/communication etc. Although these pairs of words do not explicitly overlap, they are thematically related thus conceptually similar (Lewis, et al, 2006; McCarthy et. Al, 2005).

Coh-Metrix uses three methods to compute LSA cohesion: LSA sentence adjacent, LSA sentence to all, and LSA paragraph to paragraph. The first index measures the degree to which one sentence is conceptually similar to the next sentence. LSA sentence to all calculates how one sentence is conceptually similar to every other sentence in the text. And the last method is LSA paragraph to paragraph. In this method the conceptual similarity of one paragraph to every other paragraph of the text is computed. LSA value varies between 0 and 1. A score of 0 means the absolute absence of semantic similarity and 1 signifies perfect conceptual similarity between two given text segments. In this study only the average of the three LSA indices are reported.
3.5.1.6. Situation model indices

The last class of indices deals with those elements in the text that represents the situation or mental model of the text. Situation model refers to the content or microworld that the text is all about (McNamara, Louwerse, Graesser. 2003 & 2005). For an expository text the situation model refers to a reader’s background knowledge about the text’s subject matter. For a story, it refers to the people, setting, actions, events, emotions, and various mental states of the people in the constructed microworld. In the construction of the situation or mental model the elements of the explicit text plays an important role. The interaction of the explicit text and the readers’ background knowledge help create the intended microworld of the text. To create a situation model, different elements of a text interact. Words such as connectives, particles, nouns and verbs play a major part. Verbs carry much of the content of the text and indicate different dimensions of situation models, and different forms of connectives like particles, transitional adverbs and conjunctions are used by the author or reader of a text as signals that indicate how to integrate information into the text and what should be inferred about the relation between two events (Cain. 2003; Graeser, 2004).

Coh-Metrix captures and assesses 4 of the situation model dimensions investigated so far: causal, intentional, temporal, and spatial based on the assumption that cohesion can be measured separately for each dimension. That is to say that there are different levels of cohesion on each of the dimensions (Dufty et al. 2004).

Causal dimension is the extent to which the sentences are causally related. Causal cohesion is typical of texts that refer to events and actions that are causally related such as science texts and stories that have action plots. Using the WordNet, Coh-Metrix identifies and counts the number of all the main verbs that are causal based on the word information provided by the WordNet. These particles include some conjunctions, transitional adverbs, and connectives such as so, because, as, since, consequently, etc. The more the number of causal verbs such as cause, make, kill, etc. are in the text, the more the text is assumed to convey causal content.

Another dimension captured and assessed by Coh-Metrix is **Intentional dimension**. This dimension indicates the degree to which sentences are intentionally interrelated. Again as in the assessment of causal dimension, Coh-Metrix uses WordNet data base for...
identifying and counting the intentional verbs and particles. Intentional verbs refer to the verbs or actions that are performed by animate characters pursuing some specific purposes. These verbs include *want, intend, wish, like*, etc. There are also some particles such as conjunctions, adverbs, and connectives that establish intentional relations between sentences and clauses and hence create intentional cohesion in the text.

The third situation model index provided by Coh-Metrix is *temporal dimension*. This type of cohesion is created when the text refers to event or actions. These events and actions can be expressed in different tenses (past, present, future) and their relevant aspects (progressive, completed, vs. static). The software application measures temporal cohesion by the repetition scores of certain verbs and aspects that are classified in a particular tense.

Finally, the last situation model dimension which is assessed by this software is *spatial dimension*. This is done by using the WordNet to identify and count in the text the location nouns like *university, New York, home*, etc., motion verbs such as *walk, go, run, drive*, etc. These verbs are linked by particles that refer to spatial and deictic indices such as *to, from, through, by, here, and there* (www.cohmetrix.meinphis.edu).

### 3.5.1.7. Other measures

Besides the indices used to measure the cohesion and coherence of the essays in this work, Coh-Metrix gives other measures of textual features as follows:

**General word and text information:** The third class of indices refers to the General Word and Text Information. Coh-Metrix provides general information on word and text units, i.e. the number of occurrences per 1000 words. It involves such *basic counts* as the number of words in the whole text, number of sentences, number of paragraphs, words per sentences, and sentences per paragraph.

**Readability indices:** Among almost 40 readability formulas available, this software uses two widely used ones: a) Flesch Reading Ease and b) Flesch-Kincaid Grade Level. Both of these formulas use the number of words, sentences, and syllables to estimate the difficulty or ease of the text. Flesch Reading Ease index gives a score of 0 to 100 to a given text. Higher scores indicate material that is easier to read; lower numbers mark more-difficult-to-read passages. On the other hand, Flesch Kincaid Grade Level converts
Flesch Reading Ease index to an American Grade Level. The score indicates the grade level or the number of years of education, especially if it is higher than 12.

**Frequencies:** Another type of textual information that this software deals with is the frequencies. It includes information regarding the raw frequency of content words (words that carry rich conceptual content such as verbs, nouns, adjectives, and adverbs), log frequency of content words (taking the log of the frequencies rather than the raw scores is compatible with research on reading time), minimum raw frequency of content words (the lowest frequency score among all of the content words in each sentence and a word with the lowest frequency score is the most rare word in the sentence ranging from 0-1000,000), log minimum raw frequency of content words (The word with the lowest log frequency score is the most rare word in the sentence between 0-6).

**Concreteness:** Among others, is another type of word information that this computer application measures. One way of determining the concreteness of words is drawing upon the MRC and Psycholinguistic Database which contains about 150,837 content words and gives information about 26 different linguistic characteristics of these words. Coh-Metrix measures the degree to which a given text contains concrete words. The resulting scores vary between 100 and 700. Higher scores tend towards more concrete and lower numbers lean towards more abstract words (Graeser, 2004).

Another way in which the software determines the degree of abstractness or concreteness of a word and ultimately of a text is the hypernymy and polysemy values in WordNet. Polysemy is measured as the number of meanings of a word and hypernymy value is measured as the number of levels in a conceptual hierarchy that is above or superordinate to a word. Some sets of words are functionally synonymous, for they have the same or similar meanings. The synonym sets are also related. Using the WordNet, Coh-Metrix, determines the hypernym values of words. The more hypernym levels a word has, the more concrete it tends to be, and vice versa. By the same token, the more polysemous a word is, the more it is difficult to process, especially for less skilled and less knowledgeable readers (Harabagiu, 1998; Graeser, 2004).

**Text Complexity:** The software also measures the complexity of a text. Complexity of a text involves the syntactic composition which is characterized by the frequency of certain syntactic constituents or elements in the text. For instance, a text which is
considered as syntactically complex contains considerable number of modifiers like adjectives, adverbs or determiners per noun phrase. The sentences in the text also include higher level constituents which control for the number of words. They are structurally embedded and contain a large number of verb phrases. Moreover, there are a larger number of words before the main verb of the main clause in syntactically complex sentences, so this could lead to difficulty in comprehension and putting too much demand on the reader’s memory.

*Type-token ratio:* Type-token ratio refers to the number of unique words divided by the number of all the words in the text. A *type* refers to each unique word and a *token* refers to each instance of that particular word in the text. For example, if the word internet occurs 10 times in a student’s essay, its type value is 1 and its token score is 10. In fact, this software measures the type-token ratio for content words (nouns, verbs, adjective, and adverbs) only. Type-token ratio of 1 signifies that each content word is used only once, so too many content words are introduced into the text. This could result in comprehension difficulty because there are too many new words in the text that must be processed and decoded by the reader. Smaller ratios mean that a type or unique word is repeated and fewer words are there to be decoded in a text, and thus they lead to easier comprehension.

Taking all the features of this software into consideration, Coh-Metrix was found to be the right means for assessing the features of local and global coherence of the essays objectively and with reasonable accuracy. The analysis of the essays followed a step by step and straightforward procedure:

### 3.5.1.8. Procedure

The first step in the analysis of the essays was accessing the Website of the software at (www.cohmetrix.memphis.edu). Each essay, which had previously been typed, was entered into the software through a copy-paste facility. The following information was provided for each text for later identification, retrieval and viewing of the results:

- **Title:** IR1, IR2 etc., and IN1 IN2 etc.
- **Source:** Student essays
- **User code:** IR1, IR2 etc., and IN1 IN2 etc.
- **Genre:** Science
- **LSA:** College Level
After each essay was fed into the computer application, it took only a short while for the software to analyze and return the resulting measures in the format of a Microsoft Excel data file. These measures include about 60 textual features many of which play a direct or indirect role in the coherence and cohesion of a text. Nonetheless, as noted earlier, only those features that have direct relevance to the coherence of a text and that correspond to Halliday and Hasan’s taxonomy of cohesive devices were selected for final calculations. Finally, all the data were imported from Excel to SPSS (Statistical Package of Social Sciences) for the required computations.

### 3.5.2. Topical Structure Analysis

Another instrument which was used in the assessment of the coherence of the essays written by the Iranian EFL and Indian ESL students was Lautamatti’s (1987) Topical Structure Analysis (hereafter TSA). TSA was discussed in detail in the previous chapter, review of related literature. However, for the purpose of the current section it is enough to point out that TSA works based on the assumption that in a coherent text the keywords and ideas are repeated and the sentences are arranged according to a sequence that relates all the central ideas and thoughts in a paragraph. TSA looks at ways in which the topic and subtopics can move and progress in the text in order to produce clarity and comprehensibility.

In this method, the text is divided into T-units. A T-unit is referred to by Hunt as one main clause along with all the subordinating clauses attached to it (as cited in Simpson, 2004). T-units can be considered good units for the analysis and comparing texts because they can be applied to texts that do not follow the accepted standards of writing like punctuation and capitalization.

According to Lautamatti, there are three kinds of topical progression: parallel, sequential, and extended parallel explained as follows:

a) In parallel progression, the topical subjects of two successive sentences are the same, for example:

Do you know **Internet** has uses and abuses for daily life; **Internet** is used in all over the world. (These are originally the students’ sentences. No corrections have been made.)

(Essay 3, Appendix I, A)

b) In Sequential progression, the topic of the next sentence is taken from the
c) And finally, extended parallel progression is a kind of progression which is temporarily interrupted by sequential progression.

The internet has made a tremendous impact on the lives of people worldwide. Many people are improving the quality of their lives by continuing and completing their education by taking courses online. Students, professionals, and families alike are reaping the reward of saving on day care expenses by working at home. The technology of the internet has revolutionized the way personal business is handled such as the simple task of paying bills online.

Lautamatti’s TSA describes three basic sentence elements which play an important part in her model. These elements are as follow:

1) Initial sentence element: This element refers to the discourse material that is placed at the beginning of a sentence. She identifies five types of sentence initial elements: a) connectives such as *therefore*, b) illocutionary markers like *to illustrate the point*, c) modality markers such as *the results of the study indicate*, d) metalinguistic phrases such as *it seems interesting to*, e) and finally, metatextual phrases like *next we will discuss*.

2) The mood subject: this refers to the grammatical subject of the sentence.

   *His friend* rushed him to the hospital.

3) Topical subject: This sentence element refers to the main idea which is discussed in the sentence. It may or may not be the mood or grammatical subject of the sentence, for example,

   *It was his friend* who rushed him to hospital.

In this sentence *it* takes up the position of the mood or grammatical subject while *his friend* still remains as the topical subject of this sentence.

The following paragraph selected from the corpus of the students’ essays demonstrates how topical progression is achieved. Initial sentence elements are marked by
In this modern age, the internet helps the student a lot with the advancement of technology. Nowadays internet is very popular and helpful thing for students. The first way in which the internet can affect the life of students is that it provides the vast ocean of knowledge for students. Only by clicking a button of the mouse one can get information about everything and anywhere. Now students do not have to search in library about a particular thing or topic for he is searching on the internet. Internet fulfills the needs of students as a dictionary, books, teacher and informer. Nowadays students can do a course at their home with the use of internet. Students can entertain themselves by viewing movies and listening to music on the internet.

The above paragraph consists of 8 clauses or T-units. The following progressions are identifiable in the paragraph based on Lautamatti’s topical progression:

Parallel progression: clauses 2 and 3 have the word internet and pronoun it as their topical subjects. "Internet" in the second clause is both the grammatical subject and the topical subject, so is "it" in T-unit 3.

Sequential progression: clauses 4 and 5 represent sequential progression: unlike clauses 2 and 3, the words "information" and "students" are the topical subjects of clauses 4 and 5.

Extended parallel: After a temporary break in the topical development, the word “internet” is again brought into focus as the topic of the clauses 6, 7, and 8.

Procedure

The first step in the application of TSA was coding the essays. The texts had to be coded for the identification of the three topical progression patterns which were explained earlier.

To perform the coding task, the researcher used clean computer printouts of the essays. Some essays were comparably too long, so the first two paragraphs of these essays were selected. The rationale behind this selection was that the first paragraph of an essay is normally regarded as the introduction and previews the major topical subjects to be discussed in the body of the essay. And the body paragraphs provide more development and elaboration and, hence, contain more T-units and progressions.

As the first step, the researcher identified and numbered all the T-units in the texts.
He then underlined the topical subject of each T-unit. As the next step the researcher
looked for the patterns of semantic relationship between various topical subjects. To this
end, he located all the topical subjects and plotted them on a blank page in such a way that
a visual representation of all the progressions was produced. On the leftmost part, the
clause number was indicated. The topical subject in the corresponding clause was written
across each clause number. All the words with similar or equivalent meanings were put
under the same topical depth, thus forming parallel progression or extended parallel
progression. The new topics were indented to the right.

The following essay taken from the students’ corpus (without any corrections)
illustrates the steps followed.

1) In this modern era, Internet has become the most powerful and famous
communication system. 2) It feels like the entire world is in your hand. 3) Even in
schools at a very early age the students are taught computers. 4) Even, the
engineers, doctors, businessman are not left behind in their respective fields and 5)
with the help of Internet the entire world is under time. 6) This mode is very
cheapest way to communicate to your friends when you are in abroad. 7) Well, on
the other hand there are a lot of abuses. 8) It’s leaving a worst impression among
the young generation. 9) The cyber crimes are increasing day-by-day. 10) The
social sites like orkut and hi5 are leaving a worst impact on young generation. 11)
People really get addicted to internet. 12) They sit and chat for hours and hours.
13) So, it’s better to use it in limits.

(Essay 4, Appendix I, A)

<table>
<thead>
<tr>
<th>T-Units</th>
<th>Topical subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>internet</td>
</tr>
<tr>
<td>2</td>
<td>the entire world</td>
</tr>
<tr>
<td>3</td>
<td>computers</td>
</tr>
<tr>
<td>4</td>
<td>the engineers, doctors, and businessman</td>
</tr>
<tr>
<td>5</td>
<td>Internet</td>
</tr>
<tr>
<td>6</td>
<td>this mode</td>
</tr>
<tr>
<td>7</td>
<td>abuses</td>
</tr>
<tr>
<td>8</td>
<td>worst impression</td>
</tr>
<tr>
<td>9</td>
<td>cyber crimes</td>
</tr>
<tr>
<td>10</td>
<td>the social sites like orkut and hi5</td>
</tr>
<tr>
<td>11</td>
<td>people</td>
</tr>
<tr>
<td>12</td>
<td>they</td>
</tr>
<tr>
<td>13</td>
<td>it</td>
</tr>
</tbody>
</table>
In the above representation, there are 2 parallel progressions (Clauses 5 and 6, 11 and 12), 9 sequential progressions (Clauses 1, 2, 3, 4, 7, 8, 9, 10, 11) and 2 extended parallel progressions (Clauses 1 and 5, 6 and 13).

In the meantime, the researcher also identified the number of words, number of sentences, number of T-units, and the number of error-free T-units. Moreover, to further investigate how these students preferred to carry out each progression in their essays, the researcher classified the topical subjects according to the lexical and phrasal category, i.e. Nouns, Noun phrases and Pronouns.

One main concern of the researcher regarding the analysis of the essays using Topical Structure Analysis was the subjectivity of coding task. Since the identification of topical subjects highly depends on intuition, the researcher asked another coder to code 20% of the essays. He was a teacher of English with almost 15 years of teaching experience. Prior to the coding process, he was given detailed instructions on how to go about the coding task. Correlation coefficient set the inter-coder reliability at 0.83 which was statistically significant.

After the coding was over, the researcher transferred the coding details to a tally sheet allowing one column for one specific detail. For instance, one column was given to the number of words, one column to the number of sentences, one column to the number of T-units, and so on. All this data was entered into the SPSS for the analysis process. All through the statistical analysis process the researcher enjoyed the advice of experienced teachers of statistics.

3.6. Data Analysis

In the previous sections of this chapter the researcher dealt with different aspects of data collection and methodology of the research such as the data collecting and processing procedure, selection of the subjects, details of the instruments used, and the research design and the variables involved. What follows is a description of statistical procedures used to provide a descriptive statistics of the variables and test the research hypotheses.
Statistical analysis of the data

This study is of a comparative nature. It involves a fairly large number of textual cohesion and coherence features to be compared. These features fall into two major categories: 1) those measured by the computer software, Coh-Metrix. 2) and those which were obtained as a result of the Topical Structure Analysis (TSA). Therefore, in order to test the research hypotheses and achieve the research objectives, a number of statistical tests have been considered. The students’ texts selected for the analyses varied in length and this could affect the incidence score of individual features. For instance a text of 500 words that contains 15 grammatical errors cannot be regarded poorer than a text of 150 words that contains 10 grammatical errors because the longer the text is, the more are the chances of the incidence of errors. Therefore, in order for the variables to be comparable, the length of the texts had to be normalized. To this end, the incidence of the individual features was measured based on their occurrence in every 1000 words.

After the normalization of the passage lengths, the variables were ready to be statistically processed. There were many tests at the researcher’s disposal, but each test had some underlying assumptions about the variables and the distribution of the data which restricted its use. For instance, a common statistical test is the t-test. The underpinning assumption of this test is that the data distribution is normal, the sample size is not less than 30, and the two samples come from populations with equal variances (Hinton, 2004).

These limitations made the researcher use a variety of statistical tests. Pearson product moment Correlation coefficient test was used to see if there was any relationship between two variables. Kolmogrov-Smirnov one-sample test was used to test the normality of data distributions. If the distribution proved to be normal, a t-test was used to compare two variables; otherwise, Mann-Whitney U test was used as an alternative. To see if students’ gender had any effect on the t-test result, a Univariate analysis was performed.

Besides the inferential statistics used to compare or correlate different research variables and test the hypotheses, the researcher used descriptive statistics too to try to observe and record the patterns resulting from an examination of the means and standard deviation of the variables involved. This procedure is helpful, especially when the
compared variables are not statistically significant but show underlying trends.

Moreover, during the statistical analysis of the data, a number of questions regarding some variables occurred to the researcher. For instance, *Does the students’ gender play any significant role in the results? Or, which of the coherence and cohesion features has been more frequently used in the students’ essays?* To answer such questions, the researcher conducted supplementary analyses.

To sum up, the third chapter sought to address all the issues regarding the data gathering and recording procedures. It also tried to explain the steps taken to select the subjects and the instruments used and the rational for each and every step. It also tried to explain the nature of the data, the rating and coding procedures and recording the results, and the statistical analyses used to test the research hypotheses.

The following chapter will objectively report the results of the statistical tests conducted to confirm or reject the research hypotheses (without discussing the findings of the results or drawing inferences from these results).