CHAPTER TWO: LITERATURE REVIEW

This chapter aims to review the major principles behind Error Analysis (EA) and Computer-aided Error Analysis (CEA) as these are the two major approaches in this study of learner language. The following sub-sections will also describe the various studies on learner corpora outside Malaysia, as well as in Malaysia. The final section in this chapter will re-define MWU for the purpose of the analysis of MWU errors in this study.

2.1 Error Analysis (EA)

There is much to write about on error analysis (EA) as there is abundant literature on it. For the purpose of this study, we shall keep to the fundamental understanding of 'errors', the emergence of EA, and the methodology in EA research, which is relevant to this study.

2.1.1 Definition

'Errors' is an important key word in this study, which tends to be used interchangeably with 'slips' and 'mistakes'. It is crucial to define 'errors', at the very beginning, and distinguish 'errors' from 'slips' and 'mistakes'. "Errors are the flawed side of learner speech or writing" is a simple definition given by Dulay et al. (1982: 138). Ghadessy (1980: 96) distinguishes 'errors' as "deviations which reveal the underlying knowledge of language to-date" from 'slips' and 'mistakes', which are "product of chance circumstances". On a similar note, Brown (2000: 217) also insists that "mistakes must be carefully distinguished from errors of a second language

learner". He defines an error as "a noticeable deviation from the adult grammar of a native speaker" which also reflects the competence of the learner. He continues to explain that "a mistake refers to a performance error that is either a random guess or a 'slip', in that it is a failure to utilize a known system correctly" and "mistakes, when attention is called to them, can be self-corrected". Errors, on the other hand, often indicate the learner's competence in the target language and they are 'evidence' which can reflect the learner's language proficiency. "The fact that learners do make errors, and that these errors can be observed, analysed, and classified to reveal something of the system operating within the learner, led to a surge of study of learner's error, called *error analysis*" (ibid.: 218). This is the next topic of discussion in the following sections.

2.1.2 Emergence of error analysis

Error Analysis (EA) emerged as the next paradigm to replace Contrastive Analysis (CA). CA was based on a structural approach to analyse the interference of the first language system with the second language system. The dominant belief in CA during the 40's and 50's was that a statement of the similarities and differences between various languages was enough to deal with the problem of teaching these languages (Ghadessy, 1980).

In CA, the errors made by learners are predicted by identifying the linguistic differences between their first language (L1) and the target language (TL). Interference was believed to be the main cause of error production when the learner transferred native language 'habits' into the TL. Upholding this belief, CA is deeply rooted in behaviourism and structuralism. The outcome of this is the behaviourist

theory of language which sits upon the belief that language is essentially a set of habits, whereby in the process of learning new habits, the old ones will interfere. This is called the 'mother tongue interference' (Norrish, 1983: 22). Therefore, in language classrooms, the old habits must be drilled out and the new set of responses must be learnt.

By the early 1970s, the reliability of CA was challenged. According to James (1998: 4), "many of the predictions of TL learning difficulty formulated on the basis of CA turned out to be either uninformative or inaccurate". There were information on errors which teachers already know, there were errors which were predicted but did not materialize in the learners' language, and there were occurrences of errors which were not predicted in CA. Consequently, CA gave way to EA, which provided a methodology for investigating learner language and an appropriate starting point for the study of learner language (Ellis, 1994). The procedures involved in EA research will be discussed in the next sub-section.

2.1.3 Methodology in EA research

Since the emergence of EA, it has been an important part of language pedagogy. EA became a recognized part of applied linguistics, a development that owed much to the work of Corder (1974) who suggests these steps in EA research:

1. Collection of a sample of learner language
2. Identification of errors
3. Description of errors
4. Explanation of errors
5. Evaluation of errors (in Ellis, 1994: 48)

Many studies on learner language have used these steps to analyse learner errors in the 1970s. In fact, according to Ellis (ibid.), EA was one of the first methods used to investigate learner language, which achieved considerable popularity in the 1970s, replacing contrastive analysis. More importantly, there was a boom in EA research. There are many researchers who attempted to discover more about second language learning through the study of learners' errors, especially with the desire to improve pedagogy. We shall now turn to each of the steps in EA research.

The first procedure of EA is to collect samples of learner language. The size of sample could be massive, specific or incidental. A massive sample is a collection of samples of language use from a large number of learners in order to compile a comprehensive list of errors, representative of the entire population. A specific sample consists of one sample of language use collected from a limited number of learners. An incidental sample is one sample of language use produced by a single learner.

The second step is identifying the errors. At this stage, the most crucial question which needs to be answered is 'What is an error?'. Corder (1967) distinguishes 'errors of competence' from 'mistakes in performance' and puts forth the argument that EA should investigate only errors. James (1998: 62-89) has an extensive chapter on the definition of 'error' whereby he even measures deviance (using these four categories: 'grammaticality', 'acceptability', 'correctness', and 'strangeness and felicity') and classifies them into 'slips', 'mistakes', 'errors' and 'solecisms'. Generally, most EA research will keep to a clear definition of error, such as that put forth in section 2.1.1.

In the third step – the description of errors, "one of the prime purposes of describing errors was that this procedure reveals which errors are the same and which are different, and this was a necessary step in putting them into categories" (James, ibid.: 97). The EA literature is rife with studies on the various classifications of errors. Dulay et al. (1982: 146-197) present the most useful and commonly used bases for the descriptive classification of error in these four major taxonomies: 1) Linguistic Category Taxonomy, 2) Surface Strategy Taxonomy, 3) Comparative Taxonomy, and 4) Communicative Effect Taxonomy. In their work, each of the taxonomies is described in detail based on the error types and examples of learner error. James (ibid.: 106) takes a special interest in the 'Surface Strategy Taxonomy' in his own EA research but renamed it as 'Target Modification Taxonomy'. The 'Target Modification Taxonomy' will be explained in greater detail with examples of learner errors in Chapter 4.

The fourth stage is an attempt to explain the errors based on the cause and sources of errors. By identifying the sources, it is hoped that there will be new findings which can help teachers to take another step toward understanding how the learners' cognitive and affective processes relate to the linguistic system and to formulate an integrated understanding of the process of second language learning (Brown, 2000). He has broadly categorised the sources of errors into: 'interlingual transfer', 'intralingual transfer', 'context of learning', and 'communication strategies'. (ibid.: 223-227). Very similar to Brown's, James (ibid) also has listed four main diagnosis-based categories of learner errors ('interlingual', 'intralingual', 'strategy-based', and 'induced errors'), which he expands further into various sub-categories. We shall revisit this in Chapter 5.

Finally, the fifth stage which involves the evaluation of errors, affects the learners who make the errors. The outcome of the final step should be pedagogically motivated – to create better teaching and learning materials which will help teachers to improve their teaching, as well as for learners to learn more effectively.

For two decades, EA methodology was used as a means of investigating learner language until the emergence of Computer-aided Error Analysis (CEA). After two decades, EA is considered "traditional" as the technique of Computer-aided Error Analysis (CEA) is now a new approach to the analysis of learner errors (Granger et al., 1998). In the next section, we will look at what is CEA and how CEA is different from EA.

2.2 Corpus Linguistics and Learner Corpus

The origin of Computer-aided Error Analysis (CEA) is corpus linguistics. It is necessary to provide a brief history in order to understand corpus study, define the term and describe some of the learner corpora available.

2.2.1 Corpus linguistics

Even though the term corpus linguistics first appeared only in the early 1980s, corpus-based language study has a substantial history which dates back to the pre-Chomskyan period. Instead of computers, linguist would have used shoe boxes or other storage methods, filled with papers on simple collections of written or transcribed texts. Nevertheless, the methodology was corpus-based as it was empirical and based on observed data.

The corpus methodology was severely criticized because of the 'skewedness' of corpora. In the late 1950s, the paper-based corpora were vulnerable to being skewed because it was impossible to collate and analyse large bodies of language data using papers and human hands and eyes. With the development of computer technology which offer increasing processing power and massive storage at an affordable cost, the interest in corpus methodology was rekindled.

The first modern corpus of the English language, the Brown corpus, was built in the early 1960s. The Brown corpus (i.e. the Brown University Standard Corpus of Present-Day American English) was a corpus of written American English, which was compiled using 500 chunks of approximately 2000 words of written texts. Using the same sampling techniques as the Brown corpus, the LOB corpus (Lancaster-Oslo-Bergen Corpus of British English) was created to represent written British English used in 1961. These two corpora provide an ideal basis for the comparison of the two major varieties of English as used in the early 1960s.

From the 1980s onwards, the number and size of corpora and corpus-based studies have dramatically increased and corpus methodology is currently enjoying its widespread popularity. We will look at the various learner corpora in section 2.2.2.

At this point, it is appropriate to redefine 'corpus linguistics' in today's modern context. McEnery and Wilson (2001: 2) describe 'corpus linguistics' in simple terms as "the study of language based on examples of 'real life' language use" and emphasise on corpus linguistics as a methodology rather than an aspect of language requiring explanation or description which allows us to differentiate between approaches taken to the study of language. There are many ways to define a

corpus but there is an increasing consensus that a corpus is a collection of (1) machine-readable (2) authentic texts which is (3) sampled to be (4) representative of a particular language or language variety (McEnery et al., 2006: 5).

2.2.2 Learner corpora

As mentioned in the previous section, corpus-based linguistic research has developed many types of corpora based on the purpose of the study and collection of data. An increasingly popular one is learner corpora. Learner corpora are important in the study of learner language because the data which have been collected provide empirical evidence of 'real' language used by learners. In fact, the ancestor of learner corpus can be traced back to the EA era (Granger, 2007).

However, learner corpora in those days bore little resemblance to current ones (ibid.). Learner corpora today are more than just collections of data from learners. Learner corpora are systematic computerized collections of texts produced by language learners (Nesselhauf, 2004). For Granger (2003), learner corpora is also termed as interlanguage (IL) or L2 (second language) corpora, and they are electronic collections of authentic foreign or second language data.

Learner corpora are highly useful and effective in the study and analysis of learner language because the data which have been computerized and stored electronically, allows certain programmes to provide evidence and proof that certain hypotheses we have about learner language is true. For example, the hypothesis in this study is that because Malaysian learners are not exposed to MWUs, they will have problems with MWUs in their writings. To prove this, the errors will be carefully annotated and analysed using the *WordSmith Tools*, a concordance software.

With learner corpora, many aspects can be investigated at the same time, and more general questions such as the relative frequency of different types of mistakes can be addressed (Nesselhauf, 2004). What is more important about learner corpora is that once the data is computerized, these data can be analysed with linguistic software tools, from simple ones, which search, count and display, to the most advanced ones, which provide sophisticated analyses of the data (Granger et al., 2002).

2.2.2.1 Various learner corpora worldwide

The popularity of computer learner corpus (CLC) is evident as there are more and more learner corpora being compiled. Pravec (2002) conducted a survey of learner corpora and Table 2.1 below presents the currently existing corpora with the basic information about each corpus. For the full name of each learner corpus, refer to Appendix 1.

Table 2.1

An overview of existing learner corpora

Name of Corpus	Type of Corpus	Location of Corpus	Language Background	Size of Corpus
CLC	Commercial	England	Various	>10,000,000
HKUST	Academic	University of Science & Technology, Hong Kong	Cantonese	>25,000,000
ICLE	Academic	University of Louvain-La-Neuve, Belgium	Various	>2,000,000
JEFLL	Academic	Meikai University	Japanese	>500,000
JPU	Academic	University of Pecs	Hungarian	>400,000
LLC	Commercial	England	Various	~10,000,000
MELD	Academic	Montclair State University, USA	Various	~50,000
PELCRA	Academic	University of Lodz, Poland	Polish	500,000

TSLC	Academic	Hong Kong University, Hong Kong	Cantonese	>3,000,000
USE	Academic	Uppsala University, Sweden	Swedish	~1,000,000

(Pravec, 2002: 82-83, 90)

2.2.2.2 Learner corpus in Malaysia

In Malaysia, the use and analysis of computer learner corpus (CLC) have been somewhat limited. At present, there are only three corpora – the English of Malaysian School Students Corpus (EMAS Corpus), Malaysian Corpus of Learner English (MACLE) and Corpus Archive of Learner English in Sabah-Sarawak (CALES).

The EMAS corpus consists of written and spoken data from students of three different levels: Primary 5, Form 1 and Form 4 in the Malaysian school system (Malachi et al., 2008). This untagged and unedited learner corpus was collected in 2002 and consists of close to half a million words.

The MACLE corpus is still in development and aims to be a future Malaysian sub-component for the ICLE (Botley and Dillah, 2007: 78). The idea of the MACLE project originated in Lancaster in 2001, and the research group was subsequently formed at the University of Malaya (UM) in 2002. Sample collection of written work in English of undergraduates began during the academic year 2002-3.

The CALES corpus began in 2003 and it is made up of 400,000 words of argumentative essays from students taking English proficiency courses at UiTM's Sarawak and Sabah Campuses, Universiti Malaysia Sarawak (UNIMAS) and Universiti Malaysia Sabah (UMS) (Botley and Dillah, ibid.). The CALES corpus followed as closely as possible the methodological and design principles of the

International Corpus of Learner English (ICLE) where students wrote argumentative essays under timed conditions.

There are only a handful of learner corpora in Malaysia and even with these few learner corpora, the progress in learner language research is slow. There is so much potential in learner language research and it is a pity that corpus-based research work in this area has been limited. One known recent published study using data from the EMAS corpus is on student's use of modals in narrative compositions. The study employs discourse analysis with some descriptive statistics using the concordancing programme *(MonoConc Pro 2.2)* which helped to generate statistical description that aided the analysis (Malachi et al., 2008).

2.3 Computer-aided Error Analysis (CEA)

The following sub-sections will discuss the existence of CEA and describe the various stages involved in the CEA methodology.

2.3.1 Existence of CEA

Dagneaux et al. (1998) term EA based on learner corpora "Computer-aided Error Analysis" (in Izumi et al., 2005). Botley and Dillah (2007) regards CEA as "a newer flavour of EA" and it is a newer paradigm in the research area of EA. Undoubtedly, EA research is still an important area of study and it is an improved one with the use of CEA methodology. In fact, Díaz-Negrillo and Fernández-Domínguez (2006: 84) claim that "CEA finds its origin in the methodology of EA". Even though

the basis of CEA is EA, we shall look at how CEA methodology is different from the traditional EA.

2.3.2 CEA Methodology

The technique of CEA is a new approach to the analysis of learner errors, with a hope to give new impetus to EA research (Dagneaux et al., 1998). The discussion in this section aims to describe the CEA methodology and provide examples from relevant learner corpus research.

2.3.2.1 Collection of a sample of learner language

"The starting point in EA is deciding what samples of learner language to use for the analysis and how to collect these samples" (Ellis, 1994: 49). It is important to collect well-defined samples of learner language so that clear statements can be made regarding what kinds of error the learners produce and under what conditions. In traditional EA, insufficient attention was paid to identifying and controlling the factors that might potentially influence the errors that learners produced. This is one of the limitations which was highlighted in Dagneaux et al. (1998). Traditional EA is based on heterogeneous learner data. This means that learners do not have very many similarities in their language background, proficiency level, age, etc.

In a computer learner corpus research, the presence of learners' background information is very important because it provides the researcher with the means to link the findings from the corpus research to the learners' background (Pravec, 2002). For example, in ICLE, age, sex, mother tongue background, knowledge of other

foreign languages, and the amount and/or type of practical experience in the English language are incorporated into the corpus.

2.3.2.2 Data preparation

After collecting the samples of learner language, the data has to be computerized into machine-readable format. Very often the samples collected are hand-written essays and they will be key-worded into Microsoft Word format (.doc) or Notepad format (.txt). After this process, the data is referred to as a raw corpus which is a corpus of machine-readable plain texts (written or spoken) with no extra features added (Meunier, 1998). With a raw corpus, the data is ready to be run using a wide range of linguistic software tools, or it can also be annotated, or tagged for various linguistic aspects. Corpus annotation is more often carried out on written rather than spoken data and it usually involves these processes: part-of-speech (or POS) tagging, syntactic tagging or 'parsing', semantic tagging, discoursal tagging and error tagging. For the purpose of this study, the process of error tagging will be discussed further in the section below.

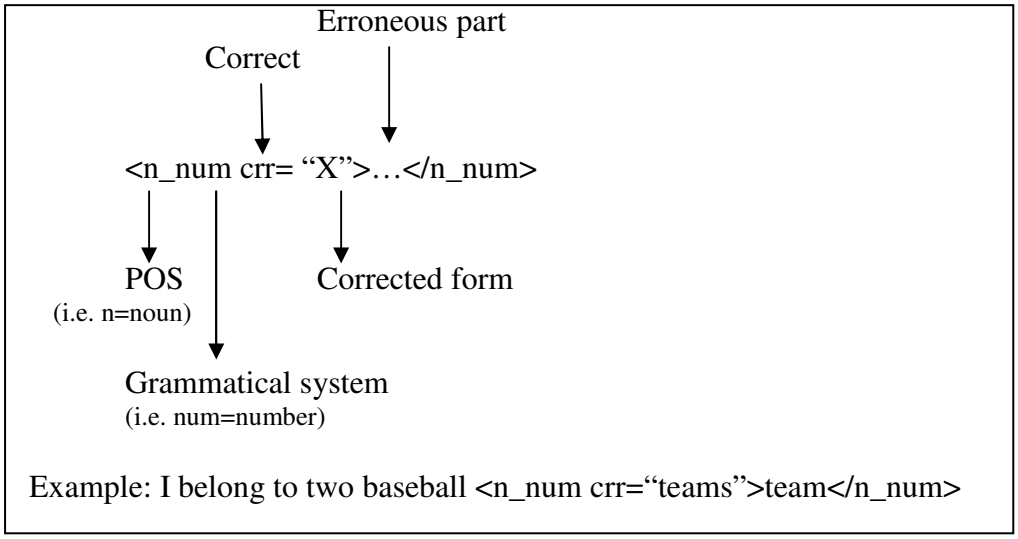
2.3.2.3 Error tagging

Error tagging is probably the most time-consuming and lengthy procedure in CEA methodology. However, once errors are fully tagged, error tags can be retrieved with the aid of software retrieval tools and analysed quantitatively and qualitatively according to the researcher's interest (Díaz-Negrillo and Fernández-Domínguez, 2006: 86). There are many learner corpora with error annotation system but not all the information is always accessible. Among the four more extensively documented error tagging systems can be found in the Cambridge Learner Corpus (CLC), the FreeText

project, the Universite Catholique de Louvain (henceforth Louvain), and the National Institute of Information and Communications Technology Japanese Learner of English (henceforth NICT JLE) (previously known as Standard Speaker Text (SST) corpus) (ibid).

Each of the error tagging system has its own tagset and how the errors are tagged depends very much on the focus of the research. Granger (2002) explains that the researcher has to make a decision whether to tag the errors in terms of their nature (grammatical, lexical, etc.) or their source (interlingual, intralingual, etc.).

For example, in the NICT JLE corpus, the original error tagset has been designed only for morphological, grammatical, and lexical errors. The error tags contain three pieces of information: Part-of-Speech (POS), morphological/grammatical/lexical rules, and a corrected form (refer to Figure 2.1 below).



(Izumi et al., 2005: 75)

Figure 2.1

Structure of an error tag and an example of an error-tagged sentence in NICT JLE Corpus

The error tagging system developed at Louvain is hierarchical whereby a series of codes from the general to the more specific is attached to each error. The first letter of the code refers to the error domain: G for grammatical, L for lexical, X for lexico-grammatical, F for formal, R for register, W for syntax and S for style. The following letter provides information on the nature of the error. For example, all the grammatical errors affecting verbs are given the GV code, which is then subdivided into GVAUX (auxiliary errors), GVM (morphological errors), GVN (number errors), GVNF (finite/non-finite errors), GVT (tense error) and GVV (voice errors). The code is tagged before each error in brackets (__) and the correction of the error is indicated with the dollar sign $__ $. Figure 2.2 is a sample of a text where the errors have been tagged using the Louvain system.

> There was a forest with dark green dense foliage and pastures where a herd of tiny (FS) braun $brown$ cows was grazing quietly, (XVPR) watching at $watching$ the toy train going past. I lay down (LS) in on the moss, among the wild flowers, and looked at the grey and green (LS) mounts $mountains$. At the top of the (LS) stiffest $steepest$ escarpments, big ruined walls stood (WM) 0 $rising$ towards the sky. I thought about the (GADJN) brutals $brutal$ barons that (GVT) lived $had lived$ in those (FS) castels $castles$. I closed my eyes and saw the troops observing (FS) eachother $each other$ with hostility from two (FS) opposit $opposite$ hills.

(Dagneaux et al., 1998: 166)

Figure 2.2

Sample of error-tagged text in Louvain Corpus

2.3.2.4 Error Analysis

After the painstaking task of error-tagging, the reward is an automated error analysis and access to detailed error statistics (Granger, 2003). Using a text retrieval software tool such as *WordSmith Tools*, it is possible to retrieve all the tagged errors according to the given tagset and sort the concordance lines in a variety of ways to bring out recurrent error patterns. For example, a search for errors bearing code XNPR, i.e. lexico-grammatical errors involving prepositions dependent on nouns, will generate all the errors which have been tagged as XNPR and list them out systematically in concordance lines as shown in Figure 2.3.

The concordance programme will also automatically generate a frequency count which indicates the number of errors for each tagset. On top of that, the concordance lines also show the corrected form which should be used in the sentence.

complemented by other	(XNPR)	approaches of $approaches to$ the subject. The written
are concerned. Yet, the	(XNPR)	aspiration to $aspiration for$ a more equitable society
can walk without paying	(XNPR)	attention of $attention to$ the (LSF) circulation $traffic$
could not decently take	(XNPR)	care for $care of$ a whole family with two half salaries
be ignored is the real	(XNPR)	dependence towards $ dependence on$ television
are trying to affirm their	(XNPR)	desire of $desire for$ recognition in our society
such as (GA) the a	(XNPR)	drop of $drop in$ meat prices. But what are these
decisions by their	(XNPR)	interest for $interest in$ politics. As a conclusion we can
hope to unearth the	(XNPR)	keys of $keys to$ our personality. But (GVT) do scientist
and (GVN) puts put	(XNPR)	limits to $limits on$ the introduction of technology in their
This dream, or rather	(XNPR)	obsession of $obsession for$ power of some leaders can

(Dagneaux et al., 1998: 168)

Figure 2.3

Sample of concordance lines – output of search for (XNPR)

This systematic analysis of learner errors is an exclusively unique technique in the CEA methodology and it is also the reason why Granger (2003: 466) describes traditional EA as "out of favour" and "gone down in history as fuzzy, unscientific, and unreliable way of approaching learner language".

In the traditional EA methodology, the extraction of errors require manual labour and this hinders the researcher from analyzing huge data as it is time-consuming and labour intensive. For example, Chan (2006) in her research, was only able to analyse 16 essays. There were eight learners and each of them contributed two essays. According to Knowles, et al. (2006), in the context of modern corpus linguistics, small amounts of data would be regarded as inadequate because it is difficult to make valid generalizations about student performance without adequate data. With CEA, a larger data can be analysed to produce more significant findings in learner language research.

2.4 Learner corpus studies

From what have been reviewed in section 2.2.2, many learner corpora already exist or have at least been started despite the fact that learner corpus compilation is a fairly new activity. According to Nesselhauf (2004), the compilation of learner corpora did not begin until the 1990s. The Hong Kong University of Science and Technology (HKUST) Learner Corpus is probably the biggest learner corpus which contains about 25 million words and it is still growing. From the survey done by Pravec (2002), there are indeed many learner corpora (refer to Table 2.1), and many studies analyzing learner corpus data are also rapidly increasing in number. However, the majority of learner corpus studies published so far have been carried out on the

basis of ICLE subcorpora, which look at advanced learner argumentative writing (Nesselhauf, 2004).

She listed the various studies on the different aspects of language which have been conducted. The major areas of language structure which have been studied to some degree are: syntax (e.g. complement clauses: Biber & Reppen 1998; tenses: Granger 1999), lexis (e.g. high-frequency verbs: Ringbom 1998), phraseology (e.g. recurrent word combinations: Milton & Freeman 1996; formulae: DeCock 1998, and discourse (e.g. connectors: Altenberg & Tapper 1998). Even though there are many studies, Nesselhauf (2004: 134) highlighted the fact that only a few of the studies have been primarily concerned with questions of second language acquisition.

With the boom of learner corpora studies, educators and language researchers are beginning to see the value of investigating learner language in second language learning. Tankó (2004) investigates the use of adverbial connectors in Hungarian university students' argumentative essays to help Hungarian writers understand the use of connectors in their writing and compares it with native speakers. The study creates awareness of the characteristics of the connectors in written English.

2.5 Defining MWUs in this study

The focus of this study is on erroneous multi-word units. As it has been briefly introduced in section 1.1.5, 'multi-word units' is a very general term and there are many sub-categories. According to Lewis (1993), the two most important groups are 'collocations', which are message-orientated, and 'institutionalised expressions', which are essentially pragmatic in character. For the purpose of this study,

'collocations' will be defined and discussed further. Bahns (1993: 57) states that 'collocation' is a term which is used and understood in many different ways. He gives a short account of how 'collocation' is understood and used by Benson, Benson, and Ilson (1986):-

> Collocations fall into two major groups: grammatical collocations and lexical collocations. Examples of grammatical collocations include: *account for*, *advantage over*, *adjacent to*, *by accident*, *to be afraid that*… They consist of a noun, an adjective, or a verb, plus a preposition or a grammatical structure such as an infinitive or clause. Lexical collocations, on the other hand, do not contain prepositions, infinitives, or clauses, but consist of various combinations of nouns, adjectives, verbs, and adverbs. Benson, Benson, and Ilson distinguish several structural types of lexical collocations: verb + noun (*inflict a wound, withdraw an offer*); adjective + noun (*a crushing defeat*); noun + verb (*blizzards rage*); noun + noun (*a pride of lions*), adverb + adjective (*deeply absorbed*), verb + adverb (*appreciate sincerely*).
>
> (ibid.: ix)

This study will focus on both grammatical collocations and lexical collocations. Due to the limitations of this study, it is not possible to discuss all the aspects of MWUs involved. Only the most revealing structures in the collected corpus will be identified for analysis and discussion. At the preliminary stage of identifying the MWU errors, these were found to be the most revealing structures in this corpus: the infinitive and modal structures (grammatical collocations), and 'adjective + noun' structures and connectors (lexical collocation). Each of these structures will be discussed further in section 3.3.

2.6 Conclusion

This chapter has discussed the EA approach as well as CEA with relation to learner corpus. At this juncture, it is important to emphasise on the importance of investigating MWU errors in learners' writing using the CEA methodology. Henry and Roseberry (2007) examine the written language of 40 Malay-speaking students in University of Brunei Darussalam using the EA approach to investigate the usage and grammar errors. What is lacking in the EA approach is a systematic methodology to identify, describe, and analyse the findings. The findings in Henry and Roseberry's study show that the errors are identified and classified. However, it lacks a systematic analysis of the errors as the errors were analysed manually.

With CEA methodology, MWU errors in learners' written language can be researched in a more empirical manner, by analyzing the actual patterns of use, with the help of a concordance programme. The CEA methodology used in this study will be discussed in further detail in Chapter 3.