CHAPTER THREE: METHODOLOGY

This study on MWU errors in learners' writing was conducted using a corpus-based methodology. This chapter will first describe how the data was collected and prepared to form a learner corpus. The explanation on the process of error identification and error tagging which follows after that will describe how MWU errors were identified and how the errors were tagged. The final subsection of this chapter includes descriptions of data processing and generation of concordance lines using the concordance software – the *WordSmith Tools (WST).*

3.1 Corpus Compilation

According to Dagneaux et al. (1998), in accordance with the general principles of corpus linguistics, learner corpora are compiled based on strict design criteria, such as the variables pertaining to the learner (age, language background, learning context, etc.), as well as the language situation (medium, task type, topic, etc.), all of which are recorded and used to compile homogeneous corpora. This is because a random collection of heterogeneous learner data does not qualify as a learner corpus. Learner corpora should be compiled according to strict design criteria and this is especially important in the case of learner data because there is so much variation in EFL/ESL and the information is not only important to the present study but also future research (Granger, 2002).

Both the learners and the task settings are important criteria and they are also specific to this learner corpus. These two aspects will be discussed further in section 3.1.1 and 3.1.2.

3.1.1 Background information of learners

In this corpus, the subjects are 17-year-old students who are in Form 5. In the Malaysian school system, this is the final year of secondary school education before they step into pre-university level. The subjects completed their six years of primary education in Chinese vernacular schools and then continued for another five years of secondary education in the same Chinese vernacular school system.

The fifth year in this secondary education is crucial for Malaysian students as they will be sitting for a very important national examination – the Certificate of Education in Malaysia, or locally known as SPM (Sijil Pelajaran Malaysia). The results obtained in the SPM examination will determine whether they qualify into tertiary institutions for further studies. One of the reasons for choosing the Form 5 students as the subjects in this study is the level of cognitive maturity of these older students. At this age, the students would be more motivated to perform well academically because they would realize that their academic performance in the SPM examination determines the path to tertiary education.

From the learners' profile, the learners all speak Chinese as their first language. The Chinese language will be used as the generic term to refer to Mandarin, the official Chinese language, and also other dialects such as Hokkien, Cantonese, Hakka, and Teochew.

The English language proficiency of the subjects in this study is average/below average form 5 students. The errors produced by these subjects are more abundant and more samples of learner language used would result in significant findings in this CEA research.

3.1.2 Task settings

According to the English language syllabus for upper secondary level (Form 4 and Form 5), it is compulsory for learners to learn five different types of essays. Table 3.1 below shows the different types of writing and an example of a title for each. The essay titles are taken from the SPM English Language examination paper, Section B: Continuous Writing (refer to Appendix 2).

Table 3.1

Essay titles

Genre	**An example of essay title**	**Number of words**
1. Narrative	Write a story beginning with: Kim was nervous when the door opened…	about 350 words
2. Descriptive	Describe an embarrassing experience in your life.	
3. Reflective	My early years.	
4. Factual	Tomorrow.	
5. Argumentative	'Teenagers today are only interested in entertainment.' Do you agree? Support your opinion.	

In this corpus, the essays written by the learners were timed and they were written in an exam-based environment. They were given one hour. As it was an exam-based writing task, learners were not allowed to refer to a dictionary, thesaurus, or any kind of English language reference materials, and no discussion was allowed too. The data were collected as exam scripts, after being graded by the teacher.

For the purpose of compiling a homogeneous data, only factual essays were collected as data in this learner corpus. The essays collected in this corpus were written based on these titles.

Essay titles:
1. Learning English is beneficial. Discuss.
2. Discuss the advantages and disadvantages of being a member of a large family.
3. A good education is the key to success.

From the essay titles, it is understood that the topic to be written on should be based on facts and the style of writing should be in a discussive manner. According to Granger (2007: 171), "the topic is also a relevant factor because it affects lexical choice, while the degree of technicality affects both the lexis and the grammar".

3.2 Preparation of Data

Handwritten essays can only exist in a computer learner corpus if it is first processed into a word format which can be read by a computer software or programme. In this study, 90 handwritten essays were word-processed into Microsoft Word. This was a meticulous task as the hand-written essays had to be typed. Every error, mistake, mispunctuation or misspelling had to be keyworded as it was. The AUTOCORRECT option within MS Word had to be switched off to prevent automatic correction of the learners' errors which could affect the analysis of the data.

Each essay was saved as a MS Word document and labeled as FS (Factual Sample), together with a number which runs from (1) to (90), which indicates the number of essays in the corpus. Appendix 3 shows the original handwritten essay of a student and Appendix 4 is the same essay which was processed into a Word document. At the end of the data collection, the total number of words is 40,000.

3.3 Error Identification

The error identification process started with the correction of the collected essays. The errors were identified and categorized into groups. The data in this learner corpus reveals various errors. However, for the purpose of this study, the investigation of learner errors is limited to four prominent structures of MWU errors: modal auxiliaries structures <MD>, infinitive forms <IN>, 'adjective + noun' collocation <JN> and connectors <CN>. In the subsections below, each of the standard structure will be explained. This knowledge is important in error identification because any deviation from the standard structure is considered as erroneous.

3.3.1 Modal auxiliaries structures (MD)

In Thomson and Martinet (1986), 'A Practical English Grammar', these are identified as modal auxiliaries: can, could, may, might, must, ought, will, would, shall and should. All these modal verbs (except ought) are followed by the bare infinitive. Any deviation from this structure is considered an error. Table 3.2 shows a few examples of erroneous modal auxiliaries structures together with the corrected form.

Table 3.2
Examples of MD errors and the corrected form

MD errors	Corrected form
can helps	can help
may affects	may affect
will not functioning	will not function
will filled	will be filled

3.3.2 Infinitive forms (IN)

The full infinitive consists of two words, *to* + *verb* (Thomson and Martinet, 1986: 212). The examples of infinitive forms are shown as below.

Table 3.3

Examples of infinitive forms

Present infinitive	To work, to do
Present continuous infinitive	To be working, to be doing
Perfect infinitive	To have worked, to have done
Perfect continuous infinitive	To have been working, to have been doing
Present infinitive passive	To be done
Perfect infinitive passive	To have been done

(Thomson and Martinet, 1986: 212)

Any deviation from these infinitive forms will be considered as errors. The data analysis in Chapter 4 will reveal the erroneous infinitive forms as well as the frequency count.

3.3.3 'Adjective + Noun' structure (JN)

Referring to Quirk et al. (1972), there are four features which are generally considered to be characteristics of adjectives. The following are the description of the four characteristics of adjectives, accompanied by an example.

1) They can freely occur in attributive position, i.e. they can premodify a noun
   (e.g. *happy* in *the happy children*)
2) They can freely occur in predicative position, i.e. they can function as subject complement, or object complement
   (e.g. *old* in *The man seemed old*)
   (e.g. *ugly* in *He thought the painting ugly*)
3) They can be premodified by the intensifier 'very'
   (e.g. *The children are very happy*.)
4) They can take comparative and superlative forms whether inflectionally or by the addition of the premodifiers 'more' and 'most'.
   (e.g. *The children are happier now. They are the happiest people I know.*)
   (e.g. *These students are more intelligent*.)

It would be too complicated to analyse all the four characteristics of adjectives in this study. The most revealing characteristic of adjectives in this learner corpus is the 'adjective + noun' structure (i.e. #1 above). The data analysis in Chapter 4 will show how this structure of MWU is a problem to Malaysian learners.

3.3.3 Connectors (CN)

The Curriculum Specifications for English Language Form 5 (Ministry of Education, 2003: 25) categorises 'connectors' into: 'conjunctions' (e.g. either…or, neither…nor, although), 'logical connectors' (e.g. furthermore) and 'sequence connectors' (e.g. later) [refer to Appendix 9(c)]. The term 'conjunctions' used in the Currriculum Specification is also known as coordinating conjunctions. The terms 'logical connectors' and 'sequence connectors' used in the Curriculum Specification are vulnerable to many interpretations because there are no specific definitions and explanation on how these two connectors should be used. Furthermore, only one example is given. In Celce-Murcia and Larsen-Freeman (1999), 'logical connectors' is a term for expressions which have been traditionally called 'subordinating

conjunctions' and 'conjunctive adverbials'. 'Conjunctive adverbials' are frequently classified according to broad discourse-functional criteria, i.e. 'additive', 'adversative', 'causal' and 'sequential'. These four broad categories were created by Halliday and Hasan (1976: 242-3). Sequence connectors are categorized as 'sequential', which is one of the four broad categories of 'conjunctive adverbials'.

In Larsen-Freeman (2000: 184), 'connectors' are categorized into 'coordinating conjunctions', 'suboordinating conjunctions' and 'sentence connectors'. 'Coordinating conjunctions' connect two similar grammatical structures, such as noun phrases, prepositional phrases or independent clauses. Common coordinating conjunctions are 'and', 'but', 'for', 'or', 'nor', 'so' and 'yet'. 'Subordinating conjunctions' connect ideas within sentences. They show the relationship between an idea in a dependent clause and an idea in an independent clause. These are some examples of 'subordinating conjunctions' according to the various categories: time (after, since, whenever), reason (because, since), result (in order that, so that), contrast (even though, whereas, although), condition (even if, unless), and location (wherever). 'Sentence connectors' show the logical connection between sentences. 'however', 'in addition', 'on the other hand', 'as a result', 'in other words', 'then', and 'later' are some of the more common examples.

Table 3.4 below illustrates how logical relationships can be formed in various sentences using the 'coordinating conjunctions', 'subordinating conjunctions' and 'sentence connectors'. Connectors are important to show logical relationships between clauses in a sentence, between sentences within a paragraph, or even between paragraphs.

Table 3.4

Types of connectors

Types of Connectors	Examples
Coordinating conjunctions	*Independent clause* Matt grew up in Kansas, **but** he now lives in San Francisco. *Independent clause*
Subordinating conjunctions	*Dependent clause* **Although** Matt grew up in Kansas, he now lives in San Francisco. *Independent clause*
Sentence connectors	*Sentence* Matt grew up in Kansas. **However**, he now lives in San Francisco. *Sentence*

(Larsen-Freeman, 2000: 184)

Based on the information given in the Curriculum Specification, and also references from the grammar books, the term 'sentence connectors' will be used to refer to both 'logical connectors' and 'sequence connectors'. Typically, 'sentence connectors' are also said to be types of cohesive devices and lexical expressions that may add little or no propositional content by themselves but that serve to specify the relationships among sentences in oral or written discourse, thereby leading the listener or reader to the feeling that the sentences 'hang together' or make sense.

The analysis of 'connector' structure in this study is concerned with the third type of connector – 'sentence connectors', which comprises of 'logical connectors' and 'sequence connectors'. In section 4.2.4, the analysis will reveal the salient

features of <CN> errors made by learners and to identify the problems learners have pertaining to 'sentence connectors'.

3.4 Tagging of Errors

With a raw learner corpus prepared, it is now ready to be tagged. In section 2.3.2.3, two examples of how errors are tagged using the NICT JLE system and Louvain system have been described. In this section, I shall describe the error tags used in this specific corpus of learner language.

The main criterion in choosing the tagset is easy recognition. For example, <MD> refers to erroneous modal auxiliaries structures and <IN> refers to erroneous infinitive structures. The preferred tag for the 'adjective + noun' structures is <JN> and <CN> is for 'connectors'. Table 3.5 below lists the error tags which are used in this study and the description for each tag.

Table 3.5
Description of error tags

Error tags	Tag description
MD	modal auxiliaries structures
IN	infinitive forms
JN	'adjective + noun' structures
CN	connectors

Figure 3.1 below illustrates how the identified errors are tagged. The samples are extracted from three different essays, as indicated by the label FS3, FS26, and FS73. Extensible Mark-up Language (henceforth XML) tags are used because

nowadays it is understood as an inherent feature of language corpora and it reputedly ensures standardization and compatibility with other systems of the same kind (Díaz-Negrillo & Fernández-Domínguez, 2006). XML is a set of rules for encoding document electronically. XML tags are preferred because the information which has been organized and described can be easily understood by humans as well as computers. The information can also be easily shared over the Internet or in other ways.

FS3
After we grow up, we need to come out for work. A good education is very important when we work. In this <JN>@competitive century/competitive society@</JN>, a person that without a good education is hard to find a job. If you want <IN>@to success/to succeed@</IN> to be a professional such as lawyer, doctor, sciencetist, a good education is the most important things that you must have.

FS26
<CN>@In opposite ways/On the other hand@</CN>, friends also will automatically help us when we are in trouble. Helping and caring each other are also the key to success. Besides a <JN>@good educated person/well-educated person@</JN> will not easily give up.

FS73
Sometimes, the lecturals of a university or college are came from other country, so without using english the students <MD>@will not able/will not be able@</MD> to know what information the lectural is giving out. <CN>@Beside that/Besides that@</CN> student <MD>@will alway using/will always be using@</MD> internet to find more information about their subject. The information in internet also mostly in english, because english is the international language.

Figure 3.1
Samples of error-tagged multi-word units (MWUs)

In this study, the opening XML tag <*> is inserted before the erroneous form and the closing XML tag </*> comes after the corrected form. The @-signs are used to indicate the position of the errors as well as the corrected form. The suggested

corrected form is separated by a forward slash (/). The learner's error is placed before the slash and the suggested correction after the slash. Figure 3.2 illustrates the error tagging system which has just been described.

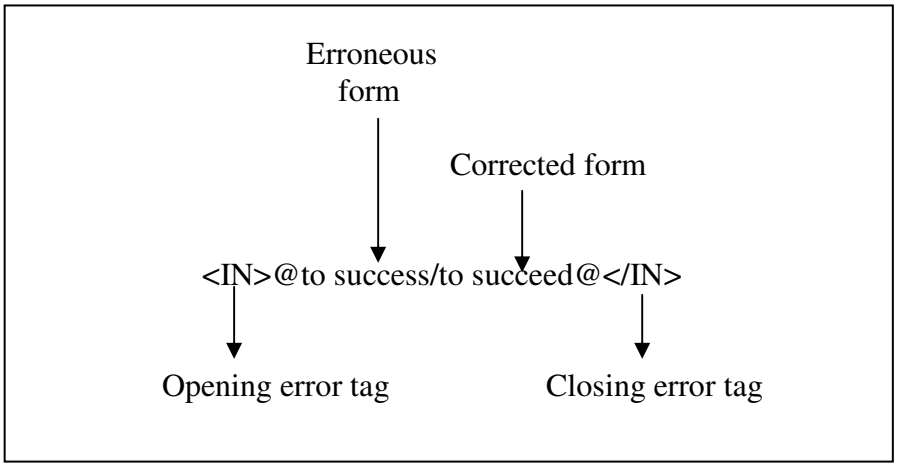


Figure 3.2
Structure of an error tag

The @-sign is used for the purpose of convenience and practicality. This system was used to tag the errors in essays written by undergraduate students in the MACLE project. This simple annotation system uses the minimum keystrokes to mark a variety of error types. Thus, it would be less time-consuming for a human researcher to insert by hand (Knowles et al., 2006).

3.5 Data Analysis

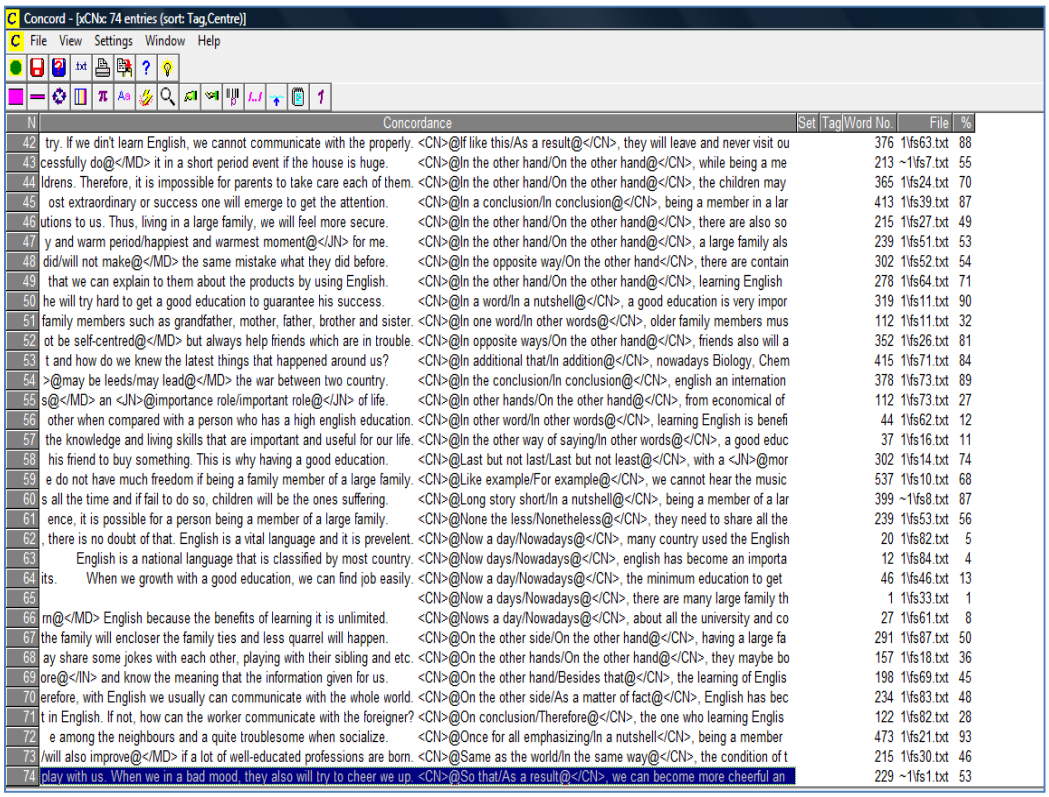
3.5.1 File conversion from .doc to .txt format

After the errors have been tagged, each essay has to be converted from the .doc format to .txt format before the data can be analysed with the *WordSmith Tools*.

It only recognizes data in .txt format because there are many hidden markups in the .doc format.

3.5.2 Using the concordance programme

The *WordSmith Tools* (WST) is used to generate the concordance lines according to the tags: <MD>, <IN>, <JN>, and <CN>. The concordance lines for the error tag <CN>, which has been called for, will be generated in a new window, as shown below.



N	Concordance	Set	Tag	Word No.	File	%
42	try. If we din't learn English, we cannot communicate with the properly.		<CN>@If like this/As a result@</CN>, they will leave and never visit ou	376	1\fs63.txt	88
43	cessfully do@</MD> it in a short period event if the house is huge.		<CN>@In the other hand/On the other hand@</CN>, while being a me	213	~1\fs7.txt	55
44	ldrens. Therefore, it is impossible for parents to take care each of them.		<CN>@In the other hand/On the other hand@</CN>, the children may	365	1\fs24.txt	70
45	ost extraordinary or success one will emerge to get the attention.		<CN>@In a conclusion/In conclusion@</CN>, being a member in a lar	413	1\fs39.txt	87
46	utions to us. Thus, living in a large family, we will feel more secure.		<CN>@In the other hand/On the other hand@</CN>, there are also so	215	1\fs27.txt	49
47	y and warm period/happiest and warmest moment@</JN> for me.		<CN>@In the other hand/On the other hand@</CN>, a large family als	239	1\fs51.txt	53
48	did/will not make@</MD> the same mistake what they did before.		<CN>@In the opposite way/On the other hand</CN>, there are contain	302	1\fs52.txt	54
49	that we can explain to them about the products by using English.		<CN>@In the other hand/On the other hand@</CN>, learning English	278	1\fs64.txt	71
50	he will try hard to get a good education to guarantee his success.		<CN>@In a word/In a nutshell@</CN>, a good education is very impor	319	1\fs11.txt	90
51	family members such as grandfather, mother, father, brother and sister.		<CN>@In one word/In other words@</CN>, older family members mus	112	1\fs11.txt	32
52	ot be self-centred@</MD> but always help friends which are in trouble.		<CN>@In opposite ways/On the other hand@</CN>, friends also will a	352	1\fs26.txt	81
53	t and how do we knew the latest things that happened around us?		<CN>@In additional that/In addition@</CN>, nowadays Biology, Chem	415	1\fs71.txt	84
54	>@may be leeds/may lead@</MD> the war between two country.		<CN>@In the conclusion/In conclusion@</CN>, english an internation	378	1\fs73.txt	89
55	s@</MD> an <JN>@importance role/important role@</JN> of life.		<CN>@In other hands/On the other hand@</CN>, from economical of	112	1\fs73.txt	27
56	other when compared with a person who has a high english education.		<CN>@In other word/In other words@</CN>, learning English is benefi	44	1\fs62.txt	12
57	the knowledge and living skills that are important and useful for our life.		<CN>@In the other way of saying/In other words@</CN>, a good educ	37	1\fs16.txt	11
58	his friend to buy something. This is why having a good education.		<CN>@Last but not last/Last but not least@</CN>, with a <JN>@mor	302	1\fs14.txt	74
59	e do not have much freedom if being a family member of a large family.		<CN>@Like example/For example@</CN>, we cannot hear the music	537	1\fs10.txt	68
60	s all the time and if fail to do so, children will be the ones suffering.		<CN>@Long story short/In a nutshell@</CN>, being a member of a lar	399	~1\fs8.txt	87
61	ence, it is possible for a person being a member of a large family.		<CN>@None the less/Nonetheless@</CN>, they need to share all the	239	1\fs53.txt	56
62	there is no doubt of that. English is a vital language and it is prevelent.		<CN>@Now a day/Nowadays@</CN>, many country used the English	20	1\fs82.txt	5
63	English is a national language that is classified by most country.		<CN>@Now days/Nowadays@</CN>, english has become an importa	12	1\fs84.txt	4
64	its. When we growth with a good education, we can find job easily.		<CN>@Now a day/Nowadays@</CN>, the minimum education to get	46	1\fs46.txt	13
65			<CN>@Now a days/Nowadays@</CN>, there are many large family th	1	1\fs33.txt	1
66	m@</MD> English because the benefits of learning it is unlimited.		<CN>@Nows a day/Nowadays@</CN>, about all the university and co	27	1\fs61.txt	8
67	the family will encloser the family ties and less quarrel will happen.		<CN>@On the other side/On the other hand@</CN>, having a large fa	291	1\fs87.txt	50
68	ay share some jokes with each other, playing with their sibling and etc.		<CN>@On the other hands/On the other hand@</CN>, they maybe bo	157	1\fs18.txt	36
69	ore@</IN> and know the meaning that the information given for us.		<CN>@On the other hand/Besides that@</CN>, the learning of Englis	198	1\fs69.txt	45
70	erefore, with English we usually can communicate with the whole world.		<CN>@On the other side/As a matter of fact@</CN>, English has bec	234	1\fs83.txt	48
71	t in English. If not, how can the worker communicate with the foreigner?		<CN>@On conclusion/Therefore@</CN>, the one who learning Englis	122	1\fs82.txt	28
72	e among the neighbours and a quite troublesome when socialize.		<CN>@Once for all emphasizing/In a nutshell</CN>, being a member	473	1\fs21.txt	93
73	/will also improve@</MD> if a lot of well-educated professions are born.		<CN>@Same as the world/In the same way@</CN>, the condition of t	215	1\fs30.txt	46
74	play with us. When we in a bad mood, they also will try to cheer we up.		<CN>@So that/As a result@</CN>, we can become more cheerful an	229	~1\fs1.txt	53

The concordance programme will list out all the errors which have been tagged <CN></CN>. A frequency count of the tagged errors is listed at the top of the window, indicated by the number of entries. Referring to the example above, there are 74 entries of the errors tagged as <CN></CN>.

The same procedure is repeated to generate all the concordance lines for the errors tagged as <MD>, <IN> and <JN>. Tagged errors which have been generated through the WST will ensure more accurate analysis as patterns are systematically organized in concordance lines.

The concordance programme also has the 'Re-sort' function which allows us to sort the evidence according to how we want to analyse them – according to the tag, centre the tagged errors, or sort the left and right environment. The 'Grow' and 'Shrink' functions are also very useful as they allow us to expand and reduce the text environment conveniently during error analysis.

3.6 Conclusion

In this chapter, we have looked at the CEA methodology which shows that the procedures involved in the collection, identification, and analysis of errors are more systematic compared to traditional EA methodology. Using the CEA methodology, the errors are identified systematically using a tagset and the tagged errors can be analysed in a more empirical manner using the *WordSmith Tools*. The concordance programme ensures accuracy and consistency during data analysis. The CEA methodology is not all perfect but at least it provides a solution to the traditional EA methodology which relies on human hand and eyes which are prone to errors.