Journal of Language and Linguistic Studies
Vol.5, No.2, October 2009

A Corpus-Based Linguistics Analysis on Written Corpus: Colligation of "TO" and "FOR"

Yunisrina Qismullah Yusuf
yunisrina@gmail.com

Abstract

This study focuses on the colligation of data in the written form. Colligation had been done to the word "TO" and "FOR" from written corpus. The research objectives were to identify the colligations of "TO" and "FOR" in their particular function as prepositions in sentences in the corpus, to discover the similarity and differences of the colligations between "TO" and "FOR" in their particular function as prepositions in sentences, and to examine whether the students had applied correctly in their essays these two particular words with reference to the function as prepositions. The data was collected from essay assignments written by Master students in the Faculty of Engineering, University Malaya. For annotation, UCREL CLAWS5 Tagset was used, with Tagset C5 to select output style of horizontal. The design of corpus used is by ICE. It was found that despite the subjects were Master students, grammatical errors were commonly found in the use of "TO" and "FOR" as prepositions.

Keywords: corpus; colligation; annotation; UCREL CLAWS5 Tagset; Tagset C5; ICE.

1. Introduction

Corpus linguistics is the study of language as expressed in real world text. In some areas it is related to computational linguistics, then at last moves towards language processing applications. This means dealing with real input data, where descriptions based on a linguist's intuition are not usually helpful.

Sinclair (1997) believed that language cannot be invented; it can only be captured. He further distinct a corpus as a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. In principle, any collection of more than one text can be called a corpus (McEnery & Wilson, 2001, p. 29). A corpus is a set of texts which is then put together

in computer-readable form for some purpose. It may consist of written texts, transcriptions of spoken material or both. The term corpora are a plural form of it.

2. Literature Review

Sinclair (1997) defined corpus linguistics as simply the study of language through corpus-based research, but it differs from traditional linguistics in its insistence on the systematic study of authentic examples of language in use.

The *American Heritage Dictionary* was the first dictionary to be compiled using corpus linguistics. It made the pioneering step of combining prescriptive elements (how language *should* be used) with descriptive information (how it actually *is* used).

Much of the materials for corpus on the Internet can be obtained free of charge. These textual materials are available for downloading. Corpora and other materials on CD-ROM or other media can cost quite a lot. Computer archives also contain large quantities of other materials which is readily available for non commercial use by academics. Some examples of English language corpora:

1. The Bank of English – written and spoken English.
2. The BNC (British National Corpus) – written and spoken British English.
3. CANCODE (Cambridge Nottingham Corpus of the Discourse of English) – spoken British English.
4. ICE (International Corpus of English) – international varieties of spoken and written English. The International Corpus of Learner' English is a part of it.
5. Brown University Corpus & LOB (Lancaster-Oslo-Bergen) Corpus – parallel corpora of written texts (but now rather outdated).
6. London-Lund Corpus (Survey of English Usage) – spoken British English (but it is now quite old).
7. Santa Barbara Corpus – spoken American English (most of the corpus is not yet available).
8. Hong Kong Corpus of Spoken English (still being compiled).
9. ICAME (International Computer Archive of Modern English) – a centre which aims to coordinate and facilitate the sharing of computer-based corpora.

10. The Association for Computational Linguistics Data Collection Initiative (USA) CD-ROM – contains *Wall Street Journal* texts, the *Collins English Dictionary* and the Penn Treebank of parsed sentences.

Some of the corpora above are available for academic use. They are of number 1, 2, 4, 5, and 10.

3. Methodology

There are many analyses that can be done by using a corpus in linguistics. It is a methodology that facilitates us in investigating language structure and use, such as to show the lexis and patterns of lexis, syntax and patterns of syntax, idiolects and speech communities, discourse and rhetoric, etc. Thus the processes of text and context retrieval sometimes provide us unexpected or unthought-of usage of language. The computer gives us the ability to comprehend data.

3.1 Scope

In relation to language studies, this present short study focuses on the colligation of data collected in August 2004. The data are in the written form. Colligation had been done to the word “TO” and “FOR” from written corpus.

3.2 Aim

To use corpus linguistics as a method in conducting this present study is as a benefit of being observable and verifiable of the naturally occurring data which had been collected. It provides systematic analysis of natural language. Before conducting this study, research objectives as guidance are as follows:

1. To identify the colligations of “TO” and “FOR” in their particular function as prepositions in sentences in the corpus.
2. To discover the similarity and differences of the colligations between “TO” and “FOR” in their particular function as prepositions in sentences.
3. To examine whether the students had applied these two particular words with reference to the function as prepositions correctly in their essays.

3.3 Corpus Description

For this present study, data were of my own gatherings which had been collected from students' essays. Many considerations were bared in mind in designing this corpus. First of all, the materials must be converted into computer-readable form. For written data, I had typed them over into the computer for hand written essays. Thus, those which were typed had been scanned by using a scanner. As this machine is not accurate 100 percent; therefore there is a need for me to perform some of editing and careful proofreading. I did not make any changes in these essays; I had typed and edited to maintain their originality.

3.3.1 Source of Corpus

Written data of this present study was collected from students of University Malaya. They are Master students in the Faculty of Engineering in their first semester. Their age ranges from 25 to 40. The class they were attending was Ergonomics Manufacturing, held in one of the faculty's laboratory every Sunday from 10:00 a.m. to 1:00 p.m. Essays assignments were collected.

3.3.2 Types of Corpus

The data for this present study were positioned in formal situations. For written data, the essays were formatted in a report form. The students of this class were divided into 5 groups to perform various measurements related to each of their topics assigned by the lecturer. They were to fill in those measurement results which were already prepared in forms. After that, they were to provide discussion and conclusion from their observation and experiment which they had conducted. The part which I had used for data were these last two parts of the essay. I did not consider the filling in the tables of measurement parts because I supposed it was not relevant for written corpus. From the discussion and conclusion parts, I was able to evaluate the content.

The standard markups applied for the corpus is by ICE (refer to Appendix 1 and 2). For written corpus, in doing wordlist and concordance, the markups had been removed to get clean information of the words used. As Nelson (1995) mentioned that in general, written texts require relatively little markups compared to spoken corpus.

Corpora may exist in two forms: unannotated or annotated. Unannotated means existing in the form of raw/plain text, whereas annotated means that words in

the text will be tagged by its proper part-of-speech, lemmatization, parsing, semantics, discourse and text linguistic annotation, phonetic transcription, prosody, or problem-oriented tagging. McEnery & Wilson (2001, pp. 32-33) explained that the utility of the corpus is increased when it has been annotated, making it no longer a body of text where linguistic information is implicitly present, but one which may be considered a repository of linguistic information. They further stated that the implicit information has been made explicit through the process of concrete annotation.

For the corpus, as I was focusing on the syntactic patterns of two of the prepositional words in my corpus, therefore it was most appropriate for me to annotate the part-of-speech of those words and other words that collocates them to grasp the implicit information. Collocation basically is a sequence of two or more consecutive words that has characteristics of a syntactic and semantic unit, and who's meaning cannot be derived directly from the meaning of its components. So this identification made recognizing the patterns of the words much easier.

For annotation, I had used the one which is already available by UCREL CLAWS5 Tagset in the Internet. To get the free CLAWS WWW trial service, I opened website http://www.comp.lancs.ac.uk/ucrel/claws/trial.html. Tagset C5 was used with selected output style of horizontal.

3.3.3 Size of Corpus

To have a representative corpus, paying attention to their size is very important. Corpus is a tedious process. It is unlikely for one person to do the work; therefore they usually work in group. However, as this was a small project, the data for this present study can be considered as a mini corpus. And the design which was used is by ICE. For this design, each file must contain 2, 000 words but not less. In corpus planning, the file must be of similar size. For this study, I had collected 10, 000 words for each form of data. Then each form of data were divided into 5 files with each file consisting of 2, 000 words.

For the data, each of the students' essays did not consisted 2, 000 words. Therefore in one file, it can consist of one or two or even until 7 essays. I had named the files based on the classification in the hierarchy of ICE Text Categories. However, I had tagged additional names at the end to confirm with the topics of the essays. The names of the files for the written corpus are:

1. **W1A-001 (AM)** – AM is abbreviation for Anthropometric Measurement
2. **W1A-002 (B)** – B is abbreviation for Biomechanics
3. **W1A-003 (SSM)** – SSM is abbreviation for Static Strength Measurement
4. **W1A-004 (MA)** – MA is abbreviation for Motion Analysis
5. **W1A-005 (PM)** – PM is abbreviation for Physiological Measurement

3.3.4 Time Period of Corpus

The data was collected in one period of time, which was in early August 2004. However, it had taken me another week for typing, scanning, editing, annotating the particular words (i.e. "TO" and "FOR") for latter syntactic patterns analysis and giving markups.

3.3.5 Programs for corpus

To study a corpus, a special analytical tool is needed. Butler (1998, pp. 217-220) mentioned some main ones which are often used:

1. Wordlist: reveals which words occur most frequently in the text(s). It will list them in descending or ascending order of frequency, or alphabetically.
2. Concordances: shows what sorts of words tend to occur in the immediate environment of a given word.
3. Distribution: shows sets of words through the various parts of the text (s).
4. Collocations: shows which particular words or sets of words enter into.
5. Keywords: a comparison with another body of text taken as a norm.

Butler further noted that in order to conduct the text analysis such as the above, there are programs which are available, such as:

1. WordSmith Tools: the most recent one, it does all the things mentioned above.
2. MicroConcord: a forerunner of WordSmith Tools, suitable not only for language researchers but also for teachers for 'data-driven learning' in the language classroom.
3. Oxford Concordance Program (OCP): very flexible but rather slow.
4. TACT: operates in two stages which is the production of database from a given text and subsequent use of the database for particular analyses without further

5. WordCruncher: consists of programs for indexing texts, and one for generating concordances. It is powerful but there is only one possible sort order, and memory limitations.

For this present study, WordSmith Tools program was used. It was available at the computer laboratory in the Faculty of Languages and Linguistics in University Malaya. In carrying out the analyses for my corpora, which were colligation and semantic prosody, the most used analytical tools from the program were wordlist and concordance. From wordlist, I could discover the words from most to least which appeared in the data. From there I chose the words which interest me most for analyses on colligation. By using the concordance analytical tool, I could observe the words or phrases that collocates the particular words which I intended to carry out analyses on.

4. Data Analysis

In written corpus, colligation is the same as syntactic patterns. Using the Wordsmith Program, I had discovered the frequency of words from the 5 files as an overall from Wordlist as much as 10, 394 of tokens (the total words in these files) and 1, 616 of types (the different words which appeared in these files). The type/token ratio was of 15.55, which means that one word was used roughly 16 times. This same word is used over and over again.

From the corpus which I had compiled, out of 1, 615 of words which the students had used, the most frequent word to occur was of course “THE”, with 820 occurrences. Other words mostly to occur (above 100) were followed by “OF”, “AND”, “TO”, “IS”, “IN”, “FOR”, “THAT”, “BE”, and “FROM.” Commencing on this frequency which was revealed by the tool, I was interested in conducting the syntactical analysis of “TO” and “FOR.” I had chosen these two words as despite of its frequent occurrences in the corpus, they are also related in some grammatical aspects which are as functions of prepositions.

5. Results and Findings

5.1 "TO"

From the concordance (refer to Appendix 1), "TO" appeared 260 times in the written corpus. With 168 concordances as infinitive marker (annotated as to_TO0) and 91 concordances as preposition (annotated as to_PRP) (refer to Appendix 3). Grammatically, "TO" can appear before a vowel and functions as a preposition and an adverb. This analysis only focused on its function as a preposition in sentences.

As a function of preposition, "TO" carries lots of positions in sentences. From these concordances, 14 colligation patterns were found for this word. In outlining these patterns, I had mentioned the word which appeared before "TO" and the phrase after it. For the column *Examples*, the numbers are the lines of where I had extracted the illustrations. They are as in the table below:

Table 1: Concordances of "TO"

No	Patterns	Concordances	Examples
1	past participle form of lexical verb + "to" + noun phrase	17	(1)…as compared to the squat lift… (79)…be assigned to strong worker… (184)…is referred to the combination... (others from data: 13, 37, 42, 44, 48, 110, 123, 124, 207, 225, 228, and 255)
2	-s form of the verb "BE" + "to" + present tense form of lexical verb + noun phrase	1	(2)…ergonomics is to specify the physical dimensions…
3	adverb + "to" + noun phrase	7	(5)…work close to the body… (82)…lie parallel to either side of… (134)…load kept close to the torso… (others from data: 81, 84, 88, and 97)
4	preposition + "to" + noun phrase	13	(10)…bending moment due to the larger turning… (153)…as opposed to using other… (204) According to Newton's first law.. (others from data: 47, 129, 142, 154, 165, 173, 182, 203, 227 and 257)
5	adjective + "to" + noun phrase	6	(24)…more prone to injury. (103)…become equal to sitting condition.

			(181)…be suitable to the parts of… (others from data: 145, 210 and 215)
6	noun + “to” + noun phrase	27	(25)…assign task to the worker… (32)…design principles to operator-hand tool… (38)…serious injury to the back… (others from data: 39, 49, 52, 61, 69, 75, 87, 107, 108, 119, 126, 135, 143, 166, 169, 171, 202, 214, 218, 222, 239, 243, 252, and 253)
7	base form of lexical verb + “to” + noun/noun phrase	5	(29)…that lead to various musculoskeletal… (105)…which contribute to the momen. (157)…is refer to local population. (others from data: 90 and 117)
8	-s form of lexical verb + cardinal number + “to” + cardinal number + noun	4	(36)…one breathes 10 to 20 times… (102)…volumes of about 60 to 80 percent… (141)…muscle after 8 to 12 weeks… (the other from data: 196)
9	-s form of lexical verb + “to” + noun	3	(64)…the table refers to each foot. (85)…which corresponds to 60 b/min… (172)…blood pressure returns to normal.
10	-ing form of lexical verb + “to” + noun	4	(149)…back amounting to hundreds of… (151)…disks, leading to increased disk… (248)……dimension following to the mean value… (the other from data: 65)
11	noun + “to” + verb + noun phrase	1	(75)…5^{th}%tile woman to a 95^{th}%tile man.
12	noun + “to” + adv + noun phrase	1	(211)…heart activity to as more amount of…
13	past participle form of	2	(95)…moment compared to sitting…

	lexical verb + “to” + -ing form of lexical verb		(249)…are required to bending more…
14	infinitive of lexical verb + “to” + noun phrase	2	(212)…can lead to discomfort or injury… (237)…can effect to the strength.

From the table above, it was found that with reference to the data, the most common syntactic patterns for “TO” were of number 1 and 6, that is **noun/verb + “to” + noun/noun phrase**. The highest rated pattern is of number 6 with 29, 67% occurrences, followed by number 1 with 18, 68%.

5.2 “FOR”

There were 148 concordances of “FOR” in the spoken corpus (refer to Appendix 2). With 10 concordances as adverbs (annotated as for_AVO21 because it is usually in a phrase of “for example” annotated as for_AVO21 example_AVO22), and 137 concordances as preposition (annotated as for_PRP) (refer to Appendix 4). Thus the analysis of “FOR” in this paper was focused on its function as a preposition in sentences.

While “FOR” functions as preposition, it holds lots of situations in sentences. A total number of 11 syntactic patterns were found for this word from its concordances. In illustrating these patterns, I had mentioned the word which appeared before “FOR” and the phrase after it. The numbers in the column *Examples* are the lines of where I pulled out the illustrations. They are as in the table below:

Table 2: Concordances of “FOR”

No	Patterns	Concordances	Examples
1	noun + “for” + noun/noun phrase	78	(2)…sustain this for 10 seconds. (6)…range of motion for the squat lift… (7) For the experiment, the result… (others from data: 1, 9, 10, 13, 15, 16, 17, 18, 22, 23, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 45, 46, 49, 50, 51, 57, 58, 60, 61, 65, 67, 70, 74, 77, 78, 79, 80, 83, 84,

			85, 86, 87, 89, 92, 93, 95, 97, , 104, 105, 107, 108, 116, 117, 118, 122, 124, 127, 129, 131, 132, 133, 137, 138, 140, 141, 142, 145, and 148)
2	adjective + “for” + -ing form of lexical verb	6	(4)…is different for standing because... (81)…want stronger for lifting or other… (100)…from 0.48s for sitting to 0.24s… (others from data: 20, 21and 82)
3	adjective + “for” + noun phrase	15	(5)…is acceptable for small, light items… (12)…more representative for our people… (54)…knee is high for stood lifting… (others from data: 47, 48, 72, 90, 113, 114, 119, 120, 123, 125, 130, and 146)
4	past participle form of lexical verb + “for” + noun phrase	9	(11)…can be used for real-world… (14)…values given for FAA Standards… (55)…be used for task requiring… (others from data: 63, 76, 99, 110, 115, and 143)
5	preposition + “for” + noun/noun phrase	4	(37) As for conclusion, squat lifting… (56)…minutes, but for lifting it was… (144) As for kinematics analysis, we… (the other from data: 52)
6	-ing form of lexical verb + “for” + noun/noun phrase	4	(44) After lifting for 2 minute the… (66) When designing for people we… (68) After resting for 4 minutes the… (the other from data: 135)
7	conjunction + “for” + noun/noun phrase	8	(53)…squat lift than for the stoop lift. (73)…detected that for the women population… (98)…repeatedly or for long hours. (others from data: 103, 106, 111, 112, and 136)
8	-s form of verb “BE” + “for” + noun	1	(59)…it is for children to first get themselves…

9	adverb + “for” noun phrase	9	(69)…where else for the women, the… (91)…that suits best for our people. (101) However, for standing, walking, and lifting… (others from data: 102, 121, and 147)
10	base form of lexical verb + “for” + noun	2	(88) Result for No.5 can be compared with… (128)…athlete can exercise for longer before…
11	adjective + “for” + -ing form of lexical verb	1	(75)…the purpose for **refer** the dimension…(an error, should be **referring**)

From the table above, it was found that with reference to the data, the most occurrences for “FOR” was of number 1 that is rated for 56, 93%. As a result, **noun + “for” + noun/noun phrase** is the common syntactic pattern for “FOR.”

5.3 Similarities of “TO” and “FOR”

After looking at the analyses conducted, below is a table of colligation patterns of similarities on the use of “TO” and “FOR” as prepositions in sentences:

Table 3: Similar Syntactical Patterns of “TO” and “FOR” in the Corpus

No	The same syntactical patterns in corpus which “TO” and “FOR” share
1	noun + “to”/”for” + noun/noun phrase
2	adjective + “to”/”for” + noun/noun phrase
3	past participle form of lexical verb + “to”/”for” + noun/noun phrase
4	preposition + “to”/”for” + noun/noun phrase
5	-ing form of lexical verb + “to”/”for” + noun/noun phrase
6	adverb + “to”/”for” + noun/noun phrase
7	base form of lexical verb + “to”/”for” + noun/noun phrase

Consequently from the table of similarity above, it can be concluded that “TO” and “FOR” are mostly followed by a noun or a noun phrase in grammatical sentences.

5.4 Differences between "TO" and "FOR"

In favor of the differences between "TO" and "FOR," I discovered that "TO" carries more syntactic patterns as a preposition compared to "FOR." With reference to the corpus, this can be seen in the table below:

Table 4: Different Syntactical Patterns of "TO" and "FOR" in the Corpus

Differences in syntactical patterns as prepositions in corpus between		
"TO"	**"FOR"**	**No of Occurrences**
-s form of the verb "BE" + "to" + present tense form of lexical verb + noun phrase	not found	1
-s form of lexical verb + cardinal number + "to" + cardinal number + noun	not found	4
noun + "to" + verb + noun phrase	not found	1
past participle of lexical verb + "to" + past tense form of lexical verb + noun phrase	not found	1
noun + "to" + adverb + noun	not found	1
infinitive of lexical verb + "to" + noun phrase	not found	2
past participle form of lexical verb + "to" + -ing form of lexical verb	not found	2
not found	adjective + "for" + -ing form of lexical verb	6
not found	conjunction + "for" + noun/noun phrase	8
not found	-s form of verb "BE" + "for" + noun	1

Based on the findings above, the most occurred concordances of the syntactical patterns will be discussed.

For the pattern of **-s form of lexical verb + cardinal number + "to" + cardinal number + noun** for "TO", examples extracted from the corpus are:

Table 5: The Pattern of -s form of lexical verb + cardinal number + "to" + cardinal number + noun

Line	The Pattern of -s form of lexical verb + cardinal number + "to" + cardinal number + noun		
36	At rest, one breathes 10	to	20 times every minute. In
102	have volumes of about 60	to	80 percent of their athletic
141	occurs in muscle after 8	to	12 weeks of endurance training.
196	than a distance of 12"	to	13" and should not be lifted more

Therefore, from the pattern above, the function of preposition here means *up till* or *until* such as *worked from nine to five*[1]. Thus, in the preposition of "FOR," this pattern does not exist.

In favor of "FOR" with the pattern of **adjective + "for" + -ing form of lexical verb**, examples extorted from the corpus are as below:

Table 6: The Pattern of adjective + "for" + -ing form of lexical verb

Line	The Pattern of adjective + "for" + -ing form of lexical verb		
4	it is a constant. It is different	for	standing because all the joint
20	me equal to sitting condition.	For	walking this was happen at
21	been more than 32 seconds.	For	the walking activity, we see
81	Its mean if we want stronger	for	Lifting or other job we must
82	value for each population.	for	designing the product to fit
100	interval decreases from 0.48s	for	Sitting to 0.24s for lifting. The

[1] The example is extracted from http://www.freedictionary .com.

Commencing on the concordances above, the –ing form of the verb functions as gerund. If we refer to the corpus, sitting, walking, lifting, squatting, and bending are names of moment of actions of people that the students must measured in achieving data to write the essays. Thus for "TO," there were also two concordances with the pattern **past participle form of lexical verb + "to" + -ing form of lexical verb** that showed the –ing form of the verb functions as gerund as well. The illustrations from the corpus are as below:

Table 7: The Pattern of past participle form of lexical verb + "to" + -ing form of lexical verb

Line	The Pattern of past participle form of lexical verb + "to" + -ing form of lexical verb		
95	and bending moment compared	to	sitting because when sitting
249	when the joints are required	to	bending more are therefore

Accordingly, these two are followed by the same form but do not occur after the same form. "FOR" occurs after an adjective, whereas "TO" occurs after past participle form of lexical verb.

At last, on behalf of the pattern **conjunction + "for" + noun/noun phrase** for "FOR", illustrations from the corpus are:

Table 8: The Pattern of conjunction + "for" + noun/noun phrase

Line	The Pattern of conjunction + "for" + noun/noun phrase		
53	for the squat lift than for	For	the stoop lift. As for the
73	It has been detected that	For	the women population of Malaysians
98	tion was done repeatedly or	For	long hours. It would be better to have
103	uire on the joints. Whereas	For	Standing and sitting position, we
106	results are all lower except	For	measurement No.12 5th and 95th
111	ent and so forth, as well as	For	basic research into the strength
112	the squat lift is faster than	For	the stoop lift. This may be because
136	ody to the steady state. But	For	lifting needs 4 minutes to recovored

From the table above, it is discovered that “FOR” can immediately appear before a conjunction when it functions as a preposition in a sentence. Thus in the corpus this case is not found for “TO” when it functions as a preposition. However I had identified that it is applied when “TO” functions as an infinitive marker TO or adverb.

5.5 Incorrect Applications of the Prepositions

From the essays, I found some grammatical errors made by the students in applying the use of “TO” and “FOR” as prepositions. I had discovered them when doing colligations for these two words. There were other errors, but I had only focus on the words that surrounds “TO” and “FOR” as prepositions. I had elaborated them by providing examples discovered from the corpus as below. I had bold the errors to get better illustrations:

Table 9: Incorrect Applications of “TO”

Line	**Incorrect Applications of “TO”**		
2	ntropometric data in ergonomics is	to	**specified** the physical dimension of
71	ermine the human body dimension	to	**designing** the product and the work
77	ns of EF=0 and EM=0 were solved	to	**obtained** the unknown forces and
90	nt population first thing is we need	to	**determined** to focus of the product
107	Example if the product is **sell**	to	the local population the dimension

Errors for number 2, 71, 77, and 90 were found after the position of “TO.” Grammatically, a verb that follows “TO” either in a sentence of present or past tense must be in the present form. So for the incorrect use above, they should be written *to specify, to design, to obtain,* and *to determine*.

For number 107, the error was found before the word “TO.” As the sentence is a conditional, therefore the verb must be in the form of past. Therefore, the correct form would be: *if the product is sold to the local population*.

Table 10: Incorrect Applications of “FOR”

Line	**Incorrect Applications of “FOR”**		
64	dard error in this sempling is large.	For	**the** example, the standard error of

75	nless we know what is the purpose	for	**refer** the dimension. Conclusion:
108	used to select subjects for fit tests.	For	**the** example: Sitting knee high for

Errors in number 64 and 108 were discovered unintentionally when I was studying the annotations of these prepositions. “FOR” there should function as an adverb for the phrase “for example.” However, because it had been added *the* after it, the tag set program had read it as a preposition. In English, there is no such phrase as “for the example.”

The last error found for “FOR” is of number 75. The correct syntactic pattern for this sentence should be **adjective + “for” + -ing form of lexical verb**. Accordingly, the correct form of the sentence would be: *unless we know what is the purpose for referring the dimension*. However, not much comment can be made on this since only one instance was found in the corpus.

6. Discussion and Conclusion

6.1 Discussion

For written corpus, I had assumed that as it was Master’s students’ essays, they would be grammatically written. However, I was surprised to find some grammatical errors which they made in their writing, particularly in the use of “TO” and “FOR” as prepositions. Referring to their level of education, I had considered these prepositions as one of the basic and simple forms of grammar, therefore, would be general to them. As I am teaching graduate students, from this situation I realize that English is not only a problem to them. It is also for people of a higher degree of education where speaking and writing in courses are all conducted in English.

Therefore language teachers in particular must take this situation into a lot of considerations, such as on how to increase students’ ability in the four main skills of language which are reading, writing, speaking and listening. To do this, of course, is not as easy as it is said. Even though today I believe language teachers are doing their best in carrying on with their duty and responsibilities, but it is a fact that a lot more efforts, patients, and cooperation are needed in improving the students’ English. Further supports in terms of financial (such as providing more attractive and effective materials, comfortable classroom environment, and training for teachers to increase their teaching ability and methods) and non-financial (such as encouragement through

mass media) from the government are greatly needed. Environment also plays a great role in enhancing the use of English. It is a fact that we are living in a multilingual and multiracial country which has led us to use English in our *rojak language* (lots of code-switching and mixing. I do not discourage this since we cannot avoid it. But it gets unintelligible when it is overused). However, maintaining the correct use of English, especially in formal settings is essential. This strategy can aid to increase the ability of its society to speak and write better English.

6.2 Conclusion: Significance of Corpus-Based Linguistics Analysis

After conducting various linguistic analyses on my corpora by using this method, I learn that why this quite new field of computer corpus research has become more and more significant is the fact that the computer is capable of processing vast amount of materials (up to hundreds of millions of words) in very short time, and with total accuracy. Thus with human, it can take many days, weeks, or even months to do. In addition, this device can produce information from texts in a form which reveals patterns that a human would probably not notice without the help of a concordance program. However, when it comes to analyzing the data, it is limited and needs human assistance.

I gained lots of advantages using this method for conducting my linguistic analysis on my corpora. The systematic information that the program had provided offers time efficiency and effectiveness in exploring my data. This method had also provided me overall accuracy that I would have never accomplished if I had done it manually.

References

Awab, S. (1999). *Multi-word Units in a Corpus –Based Study of Memoranda of Understanding: Modal Multi-word Units.* Unpublished Ph.D. Thesis. Lancaster University.

Butler, C. (1998). "Using Computers to Study Texts." In A. Wray, K. Trott & A. Bloomer, *Projects in Linguistics: A Practical Guide to Researching Language (pp. 213-223).* London: Arnold Publishers.

Computers and Corpora. (2004). Retrieved September 17, 2004 from the World Wide Web: http://users.ox.ac.uk/~lou/wip/corpora.html

Corpora Computer Lab. (2004). Retrieved September 17, 2004 from the World Wide Web: http://www.uni-giessen.de/~ga1007/ComputerLab/corpora.htm

Corpus Linguistics. (2004). Retrieved September 17, 2004 from the World Wide Web: http://www.all-science-fair-projects.com/science_fair_projects_encyclopedia/Corpus_lingusitics

Corpus Linguistics. (2004). Retrieved September 17, 2004 from the World Wide Web: http://www.engl.polyu.edu.hk/corpuslinguist/corpus.htm

Hanston, S. (2002). *Corpora in Applied Linguistics.* Cambridge University Press.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Longman.

Kirk, J. M. (2000). *Corpora Galore: Analyses and Techniques in Describing English.* Amsterdam: Rudopi.

McEnery, T., and Wilson. (2001). *Corpus Linguistics.* Edinburgh: Edinbugh University Press.

Nelson, G. (1995). "The International Corpus of English: Markup and Transcription." In G. Leech, G. Myers and J. Thomas (eds.), *Spoken English on Computer: Transcription, mark-up and application (pp. 220-223).* London: Longman.

Sinclair, J. (1997). "Corpus Evidence in Language Description". In A. Wichmann, et al (Eds.), *Teaching and Language Corpora. (pp. 27-39).* Harlow: Longman.

Stubbs, M. (2001). Words & Phrases: Corpus Studies of Lexical Semantics. Oxford, UK: Blackwell.

Yunisrina Qismullah Yusuf was a graduate of Syiah Kuala University, Banda Aceh, Indonesia and is currently a lecturer at the English Department in the Faculty of Teacher Training and Education, Syiah Kuala University, Banda Aceh, Indonesia. She received her MA degree in Linguistics from University of Malaya, Kuala Lumpur, Malaysia. She is now pursuing her PhD (Linguistics) also in the University of Malaya, Kuala Lumpur, Malaysia. Her research interests are phonology, syntax, second language acquisition, corpus linguistics, sociolinguistics, language and culture.