CHAPTER II

LITERATURE REVIEW

2.1 Recent developments in Corpus Linguistics

Corpus linguistics is known as the study of language based on example of real life language use (McEnery & Wilson, 1996, p. 1). The word *corpus* is from the Latin word for *body*; the plural is *corpora.* Nowadays Corpus linguistics is associated with computer-assisted. It is supported by Leech (1992, p. 106) who argued "corpus linguistics is now identical with computer corpus linguistics". Computer is known as the tool that can offer the ability to store large amount of data, statistical and accurate reliability which is appropiate in corpus linguistics.

Because of trust on large scale data, the term *corpus linguistics* is related with empirical research which could be done by collecting any spoken and written text (McEnery & Wilson, 2001, p. 29). It shows that corpus linguistics analyses and describes rightly language in text. The collection of texts in an electronic database can be called as the notion of corpus. Since the development of technology, computer is used to store and search the corpus.

Brown University Standard Corpus of Present-Day American English was the pioneer on machine-readable of English Language Corpora for linguistics research created by Francis and Kuˇcera (Kennedy, 1998, p. 23). This corpus is commonly called as *Brown Corpus*. It qualifies a corpus because *Brown Corpus* contains a body of text which has one million words of edited written American English. The *Brown Corpus* is called as a "balanced "corpus because it consist of

500 samples each of around 2,000-words which represents different types (or genres) of written English both informative and imaginative prose, including press reportage, editorials, government documents, technical writing, and fiction. The structure of *Brown Corpus* relied on selection of text categories to represent a wide range of stylistic aspects of written American English. It was set as a standard corpus-based research (Kennedy, 1998, p. 27). This corpus is used for most value to individuals whose interests are primarily linguistics and who want to use a corpus for purposes of linguistic description and analysis. Many English departments all over the world have used this corpus as a source of data or the way of investigating in which computer is used in language research. This success led the evolution of text corpora on the recent events: the larger 'super corpora' of the 1980s and 1990s, and the current and the future 'cyber corpora'.

The example of super corpora which led to small and standard corpora was Birmingham corpus (1980-1986). This corpus was created to represent the English language as it was relevant to the needs of learners, teachers and other users (Renouf, 2007, p. 33). Related to the teachers and leaners, corpus linguistics shows that this study is able to contribute in language teaching. 'Research on native-speaker corpora has yielded more accurate and detailed description of English'(McEnery & Gabrielatos, 2006, p. 50). The opinion indicated that corpus linguistics is helpful for teachers and learners to find various references book for teaching and learning English language that draw on the findings corpus-based research. The first mega corpus which comprised 500 million words of British and American English was launched at the University of Birmingham. This corpus

covered both spoken and written text (McEnery & Gabrielatos, 2006, p. 48). The same period, British also built corpus project.

British National Corpus (BNC) was created by a consortium of academic industry such as Oxford University Press, Oxford University Computing Services, Longman Group Ltd, the unit for computer research on the English Language from Lancaster University, British library research and development department. This corpus covers 100 million words. Regarding the size of the amount words, British National corpus has smaller than Bank of English. However, British National Corpus has been more frequently accessed by many universities and researchers because this corpus has been freely accessible for general public.

Corpus linguists compiled the texts from different period of development of English. They believe that a language change over time to time is able to investigate (McEnery & Gabrielatos, 2006, p. 54). Consequently, they designed diachronic corpora, *Helsinki Corpus of English Text*. This corpus aims to study language change over time (Kennedy, 1998, p. 38). The corpus consists of 400 samples of text. Besides, it comprises 1.5 million words which cover the period from Old English to Early Modern English (c. 750 to c. 1700). This study is to compare text across time-frame (Renouf, 2007, p. 36). The project of corpus has remained in a progress of technology development.

In the 1990s, the web existed to store textual and other information. World Wide Web treated text as an on-line corpus which is called as Cyber-corpus. Corpora are limited in size because of the time-consuming and out-of date by the time completion (Renouf, 2007, p. 42). Web texts, on the other hand, are

available, vast in number and volume, constantly updated and full of the latest language use. Web is appropiate as a source of language data which is in big scale, free and instantly available (Kilgariff & Grefenstette, 2003, p. 333). When the Web is known and used by people, web text provides retrieving instances of words and phrases in text that are either too rare or too recent to appear in conventional text corpora. The merits of web text are freely available, vast in number and volume, updated language use (Renouf, 2007, p. 42).

The present of ukWaC (UK Web as a corpus) introduces a very large corpus of English which comprises more than 2 million words. ukWaC was constructed in 2007 by web crawling which partly funded by University of Bologna and as part of WaCky project(Web *as* copus *k*ool *y*nitiative). The designing of ukWaC is a consortium of researchers who interested in the exploration of the web as the source of linguistic data. This corpus includes both of 'pre-web' texts of a varied nature that can also be found in electronic format and texts representing web-based genres such as personal pages, blogs, posing in forum (Santini and Sharoff, 2007). The objective of ukWaC is to build a corpus of British English. This corpus is one of the largest freely available linguistics resources for English. From those developments of corpora, it can be categorized as *general corpora* which are designed in a large scale of millions of words with text collected from a wide range of sources. They act as a reference corpus.

This study used the similar principle like those corpus project by compiling the collection of written text. The corpora of this study were written text in magazine which consisted of Men's Health Magazine and Women's Health

Magazine. Its size was in small scale. The corpora were taken from the web text. Since the corpora of this study were in written text, this study focused on vocabulary. Therefore, the discussion of classification of vocabulary is given in the next section.

2.2 Classifications of Vocabulary

Because of this study analysed typical vocabulary in special text, it is important for learners who study specific purpose to know the typical vocabulary that occur. Chung and Nation (2003, p. 252) argued that the best way to determine typical vocabulary for any words is to use a rating scale that classifies words according to how closely related they are to particular subject area. Consequently, it is better to know further classification of vocabulary in text. Nation (2001, p. 104) divides the classification of vocabulary into four: high frequency word; academic vocabulary; low frequency words; and technical vocabulary. One way of identifying the classification of vocabulary is to count every word form in a written text which is called *tokens* and sometimes 'running words'. It means that every the same word form occurs more than once, then the occurances are counted.

High-frequency word covers 80% of running words of academic texts and newspapers, and around 90% conversation and novel. The notion of high frequency words are content words. The percentage indicates that high-frequency word comprise large proportion of words both in spoken and written texts. Academic words can be found in various academic texts. It comprises 8.5% of

academic text, 4% of newspaper and less than 2% of the running words of novel. This vocabulary is frequent to a wide range of academic field. The other classification of vocabulary is low frequency words which cover 5% words of academic text. This vocabulary covers all the words that are not in high-frequency words, academic words, and technical words for certain subject. Low-frequency words rarely find in our use of the language. Technical words are mostly used by people who have interest in examining in a specialized text (Chung & Nation, 2003, p. 104). The words are typically related to the topic and subject area of the text. The technical words cover 5% of the running words in a specialized text. The number of percentage indicates that the technical words are occurred frequently in a specialized text or technical corpus however they have low frequency in other fields.

Those classifications of vocabulary are the basis in identifying the typical vocabulary which can be applied for computer program. Basically, the program runs to compare the number of occurences of a word in target corpus with the number of occurences of a word in reference corpus. Furhermore, the typical vocabulary shows the use of language regarding gender. The description of language and gender is given in the following sub-sections.

2.3 Language and gender

In the terms of language and gender research, gender and sex show difference sense. *Gender* is considered as the basis of socio-culture behaviour, while *sex* refers to show a biological distinction (Holmes, 2001, p. 150). However,

many researchers prefer used the term *gender* compared to *sex*. Sociolinguists have been interested in investigating whether gender could influence the linguistic features, for instance pronunciation, grammar, and communication style.

Through her work *Language and Women's Place*, Lakoff proposed that women's language tends to used *hedge* which reflects uncertainty, *'empty' adjective* like *divine, charming etc,* which represent of feminity (Lakoff, 2004, pp. 78-79). Besides, women are diagnosed with depression more often than men. It indicates that women's language is considered as a weak and unassertive. Thus, women tend to show the powerless language. In that case, if women act like men they could be having power.

The difference in the context of language and gender can be found in linguistic marker. It mostly reveals in the usage of suffixes to word. It represents where women involve "belong" occupationally. For instance, a waiter becomes a waitresses and actor becomes actress. Furthermore, men and women are viewed to have linguistic varieties in the terms of vocabulary they used. It indicates that the vocabulary they used occur as a social phenomenon which is closely related to social attitude. It is clear that men and women are not only different biologically but also different social roles in their society. The different social roles among men and women are appeared from the roles as father and mother. In household, father is commonly expected as a breadwinner, disciplinarian, and ultimate decision-marker. The term of "woman "indicates the existence of euphemistic terms for woman's principle, that of "housewife" (Lakoff, 2004, p. 52). The roles as a housewife and a mother have closely roles of love, nursing, and self-sacrifice.

Another factor may influence the using of language between men and women are psychological intervention. The value of psychological intervention could occur in the using of particles which comprises pronoun, preposition, and auxiliary verb. Those features can give sense of noun and verb as marker of emotional state, social identity, and cognitive style (Pennebaker, Mehl, Niedehoffer, 2003, p. 547). Women more likely use longer sentence than men. it shows that women use more detail to describe, for instance events, person, place, and thing.

The word *I, you* and *we* become linguistics marker of pronoun which shows different perspectives on the relationship between speaker/author and addressee. The use of first-person singular pronoun is associated with depression and self-focus. First-person plural reflects the sense of group identity and a sign of emotional distancing (Newman, 2008, p. 216).

Women frequently use modal auxiliary verb such as *could, would,* and *may* than men. In addition, women also more likely used intensifier such as *so, just,* and *quite*. Those intensifiers are characteristics of women language than of man's. Women are very emotional. In written language, women tend to use assertive function to show their feeling. The terms of endearment such as *sweetie, dear, honey* is more likely used by women.

Lakoff found that the differences of men's and women's language occur in their word choices. In the terms of colour, women closely used *beige* and *lavender*. Besides, women tend to use *delicious* and *delicate* which refers to cooking word (Lakoff, 2004, p. 43). In contrast, men tend to use vocabulary

related to sport and auto mechanics word which indicates the masculinity. Herring (1996, pp 207-208) also supported that men tend to be more concern about threats. It shows that the sense of masculinity is typically related to the notion of men power.

In working class communities in the west, men and women show the different linguistics behaviour. Women commonly used the standard language than men (Hellinger & Bussmann, 2001, p. 251). The factor of awareness and society traits leads women to use more standard form of English. Regarding of awareness, women are more aware of their low status in society. By using standard language, women tend to show their equality. In society, women are designed as a correct model in having behaviour. It leads women to use language correctly and polite forms of language to show the using of standard language. As a consequently, women claims that they more having high status within their society.

Those principles of language and gender can be investigated using corpus approaches. The notion of corpora approaches consists of the texts that are represented in written form as word. The corpora are usually known as representative of a language variety (Baker, 2010, p. 6). Therefore, this study chose magazine as a media which consist of written text. Besides, written text has become interest in language and gender study. It led this study to investigate language and gender in written text like in magazine.

In magazine industry, the differences of language can be seen based on the target market either men or women. It is related to this research which focused on

the word choices that typically used in men's and women's magazines. In modern era, the magazines are available in electronic media. This study aims to analyse the typical vocabulary is used between men's and women's magazines by using electronic program which is able to extract the terms. The notion of this electronic program is discussed in the next sub-section.

2.4 TermoStat

In extracting typical vocabulary used in Men's Health Magazine and Women's Health Magazine, this study used computer program for analysing corpus. The previous definition of typical vocabulary is closest with Technical vocabulary. Technical vocabulary involves words that typical to the subject area to the text (Chung & Nation, 2003, p. 104). The program which can identify the typical vocabulary and run for free is *TermoStat.* It was designed to extract terms automatically in some online version such as French, Spanish, Italian, and English. However, this studies only concerned on English corpora which can be accessed at http://olst.ling.umontreal.ca/wdrouinp/termostat_web/interfaceTermostat/php.

The merits of *TermoStat* are to extract single word units as well as multi-word units, and to identify word classes such as verb, noun, adjective, and adverb. *TermoStat* is able to identify each class and all categories as well. Chung and Nation (2003, p. 258) mentions that *TermoStat* typically uses statistical approach to compare the occurrences in a technical corpus with its occurrences in a general corpus. The program analyses the typical vocabulary which has the highest

keyness in each group of magazine. “Keyness are not generally or simply key in a given language, but they may be key in a given text” (Scott, 2010, p. 43).

In determining better keyness, this study used log likelihood calculation. The Log Likelihood does not depend so critically on assumptions of normality so it can work well with small volume of text (Dunning, 1993, p. 61). It compares word frequency between target corpus and reference corpus in database of the corpus program which must have significantly frequency than expected (Dunning, 1993, p. 70). The references corpus is non-technical corpus. The target corpus is technical corpus. The Reference corpus covers 7 400 000 words from 13 746 articles. The articles were taken from *the gazette,* a Montreal-based newspaper. Besides, it ensures that the articles comprise various subjects (Drouin, 2003). Those two corpora are quite different in terms of content, so that it is possible to identify the typical vocabulary in target corpus.

2.5 Related studies

There are some researchers who have conducted research similar to this study; Ratnawati (2005) and Arum (2011). In her thesis, Ratnawati investigated in distinguish between men and women in swear words. The Swear-words used by Jimmy and Alex to their male and female addressees in the "8 Mile" movie. She used Coates’s and Holmes’s theory in supporting her analysis men and women language related to swear word. She found that men and women utter different swear words. Ratnawati (2005) stated for the final result in investigating men and women language regard swear word is that “male character used swear-words

more than the female character, especially when the male character talks to his male addressee (male to male). The female character uses more swear words when she talks to her male addressee (female to male) than to her female addressee (female to female)".

The other study conducted by Arum (2011) who studied on "a diachronic Corpus Based Analysis of the adjectival collocates of [Man] and [woman] in American English from 1861 to 2010. She used secondary data both in She compare the changes in adjectival collocates between men and women. She found that American society has changed their perception about men and women time to time.

Compared with this study, the writer found the different limitation of research both of Arum and Ratnawati. Arum focused on adjectival collocates of men and women in America in period of time. While, Ratnawati focused on the different choice of swear words which is used by men and women. This study focuses on the typical vocabulary of heath news which occurs in Men's Health magazine and women's Health magazine.