What Word Concordances Offer to Foreign Language Teachers

Ari Purnawan

Universitas Negeri Yogyakarta

(Jurusan Pendidikan Bahasa Inggris FBS UNY, No HP 0815-6867-906, e-mail aripurnawan_uny@yahoo.com)

Abstrak

Selama ini analisis *corpus* dan mesin *concordance* sudah banyak digunakan di negara-negara pengguna bahasa Inggris, namun di Indonesia belum banyak dikenal atau dimanfaatkan untuk tujuan pembelajaran bahasa dan penelitian linguistik. Artikel ini menyajikan berbagai kelebihan yang ditawarkan analisis *concordance* dalam belajar bahasa Inggris atau bahasa asing lainnya, terutama bagi pembelajar orang Indonesia. Manfaat yang dapat diambil dari analisis ini adalah kita dapat belajar bagaimana sebuah kata digunakan, membandingkan penggunaan kata tersebut oleh para penutur aslinya dan penggunaan kita, sehingga dapat kita ketahui apakah kita sudah menggunakannya secara alamiah dan tepat seperti mereka. Belajar struktur kalimat, tata bahasa, frasa, kolokasi, dan diksi menjadi lebih mudah dan dengan ketepatan tinggi karena rujukan yang diambil adalah penggunaan bahasa yang nyata. Untuk para pengembang bahan pelajaran, manfaat yang dapat diambil adalah tersedianya data untuk pemilihan kata yang sesuai, daftar kata paling sering dipakai oleh para pengguna, dan cara penggunaannya. Masalah dan tantangan bagi pengajar dan pembelajar di Indonesia adalah belum tersedianya *corpus* bahasa Indonesia yang berukuran besar yang dapat dijadikan rujukan untuk melakukan analisis seperti yang sudah terbangun pada bahasa Inggris dan bahasa-bahasa besar dunia lainnya.

Kata kunci: *corpus analysis*, *concordance*

1. Introduction

Recently the use of computer programs in education is growing very rapidly. Many educational objectives seem to be much easier to achieve after the involvement of certain computer programs. The programs offer tons of benefits, and since then computers cannot be separated from efforts to improve educational sectors. The world of foreign language

teaching is not the exception. One of the advantages that language teachers can get from such programs is that they can analyze how a word has been used: in what context or situation a word is most appropriately used, what other words collocate with it, how many times a certain word occurs in a text of 1000 words, and the like.

The forms and functions of the computer software are very numerous. A word concordance analyzer, for example, can do many different actions for different purposes. Analyzing a small number of words taken from a group of students' writing can be useful for revising their writing draft, analyzing a model text written by a native speaker may result in a fruitful discussion about how the language has been used, and analyzing a very large corpus might be beneficial for book writers, curriculum developers, linguists, or language researchers.

The use of English corpora for various purposes has been a long tradition in most western countries. Any development or advancement in scientific works concerning the use of words involves computer processing and analysis. Entries in new dictionaries, words used for books for children or for immigrants, or the list of new vocabulary to be learned by elementary school children all refer to how the language has been empirically used in the real world, and the computer software gives quick and precise solutions to them.

In Indonesia, in contrast, the use of such software for educational purposes is very limited. English textbook, bilingual dictionary, and language curriculum developers will intuitively select and use words for their products, while actually computer software can help them become more selective in deciding what to be included in their works. However, they are not the ones to blame, because the large collection of Indonesian texts stored in a digital-electronic storage is not available. Even when a book writer, for example, wants to use a collection of Indonesian texts for their reference, s/he must do it by him/herself, and therefore the number of texts that s/he can collect is far from being ideal for a source. What happens to language teachers is also the same. Teachers are not accustomed to using simple software for analyzing texts for even a very simple purpose such as analyzing the students' production.

This article aims to explain strengths that a concordance analysis offers. Many parties should be aware that the data resulting from such software can provide us with a wider perspective on how the language has been used by all people or language users and how it should be used by language learners.

2. Corpus Analysis and Concordances

The word corpus means a collection of texts, either spoken or written, that is stored in a computer database. Mohammadi (2007) states that the collection of language use in a corpus, although does not offer anything new about a language, gives a new perspective to linguistic researchers, language educators, and translators or interpreters. From the patterns resulting from a simple corpus analysis, for example, we can learn many things about a language, the language use, or a certain linguistic item of the language.

Corpora come in various sizes or numbers of entries. A mini corpus can contain only a short text with only fifty or sixty words, but a big one can have millions of words taken from thousands of different texts from different sources. The Cambridge International Corpus, for example, has over 700 million entries from various text sources. Cobuild Dictionary used a corpus of 200 million words as its source. Large-sized corpora are very useful for linguists and language teachers or learners. The corpora provide them with information of how a language or a word has been used by its speakers, how a language has changed, or how speakers use different patterns for different situations. The use of corpora for instructional purposes is also growing (McCarthy, 2002).

One of the important contributions of a linguistic corpus is that it becomes the data source of our concordance analysis. A very simple concordance that has been widely used is the internet search engine, such as the Google or yahoo search. By typing one or two key words, Google will show thousands of fragments or phrases containing the key words. A computer user then can easily find sites that have the key words in their texts. However, Google's concordance does not directly give useful information concerning the language and its use or other linguistic aspects. Most Google users have used it for searching websites. Once the search engine displays the sites, the user will soon click one of them and forget the engine. Lamy and Mortensen (2009) underline a practical use of a web search engine, for example for studying a simple collocation. If we want to know whether the phrase "a metal wood" really exists in English, we can simply type the key words in the engine. The engine gives examples of the phrase and its context, that is, how people in the sites have used the phrase in their texts.

However, for a more specific purpose, such as a collocation analysis or other word studies, Google's engine is obviously inadequate. A more sophisticated software is needed. This software is usually known as a concordancer. Several concordancers are available freely in the internet, and several others, which offer a more thorough investigation, need to be purchased if we want to use them.

One free-software that we can use is the Textanz. This software offers useful information concerning basic statistics about the text that we input, including the number of words, sentences, lines, the average words per sentence and sentences per paragraph, the longest and shortest words and sentences, and the readability level of the text. Other important items that it provides are occurrence frequency of all words contained in the text, and of course, the word concordance. Using this simple concordance, we can learn how a certain word is used in the text. Figure 1 shows how we can learn from a text about how the word 'strategy' has been used by the writer. The source text is a 200 word scientific paper.

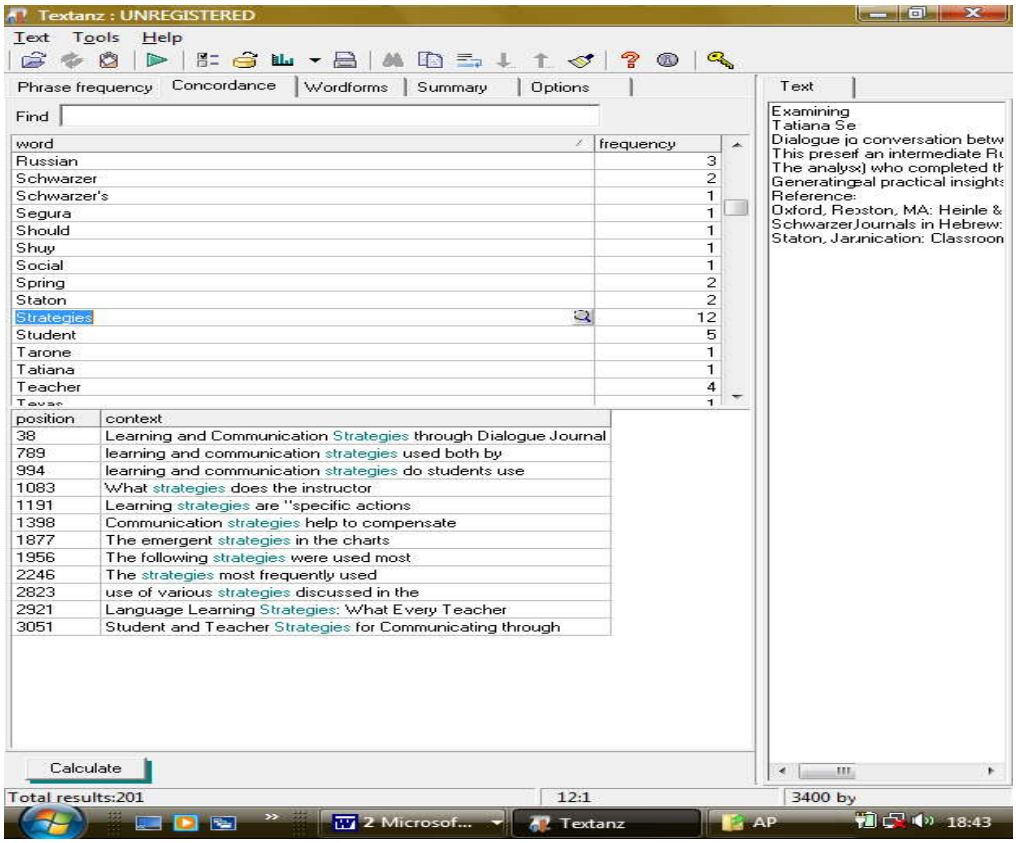


Figure 1. The Textanz program showing how the word 'strategy' has been used in a text

To use the Textanz, we can download the trial version program from the internet and we can start our own project. In the example in Figure 1, a research paper text was the source. Click the text bar, and insert the copied file into the space on the right hand side of the program page. Soon the basic properties of the inserted text appear on the left, counting the list of words used in the text and their occurrence frequency. We can choose one word from the list, and the simple concordance will appear at the bottom. In the example, the word 'strategy' was chosen, and the program shows how the word has been used by the writer. Other options are also available, such as phrase frequency, in which we can display a certain phrase and ask the program to show how the phrase was used. The word form will tell the

basic properties of a word, such as its frequency, length, and dispersion. This program is a handy tool for examining a short text, for example our own essay, and for comparing it to a more 'standard' piece of writing.

Several other frequently-used free software programs are the TextSTAT, the Concordance, and the Wordsmith. The TextSTAT software gives us a clearer concordance list with the keyword put in the centre of each fragment. Using this concordance, we can identify some pattern of how a word is used, or what words collocate with the keyword. The user can trace how the word 'appoint', for instance, has been used by people. The software will inform us the prepositions that usually follow the keyword, the nouns that accompany it, and other information concerning its use and context. With the Concordance evaluation copy program, which is also free, we can analyze a word and its context. A simple study on word frequency, occurrence, minimal pairs, collocation, and concordance can be easily conducted using one of the above programs.

If the internet connection is a problem, the programs have good news: they can be used off-line. We do not have to be connected to the internet. In remote areas where the connection to the internet is not possible, we still can conduct our word study using the programs.

A larger, professional corpus can be used for a more useful output. The Cambridge International Corpus (CIC), for example, provides a huge number of text collections which can give an illustration of how the language has been used by its users all over the world. Below is a list of most frequently used English words from a corpus of 10 millions words.

Table 1: Most frequently-used words from 10-million-word Cambridge International Corpus

Rank	Word	Frequency
1	the	439,723
2	and	256,879
3	to	230,431
4	a	210,178
5	of	194,659
6	I	192,961
7	you	164,021
8	it	150,707
9	in	142,812
10	that	124,250
11	was	107,245
12	yeah	86,092
13	he	78,932

14	is	75,687
15	on	71,797
16	for	69,392
17	but	64,561
18	she	61,406
19	they	58,021
20	have	55,892

Source: O'Keeffe, et al. (2007:35)

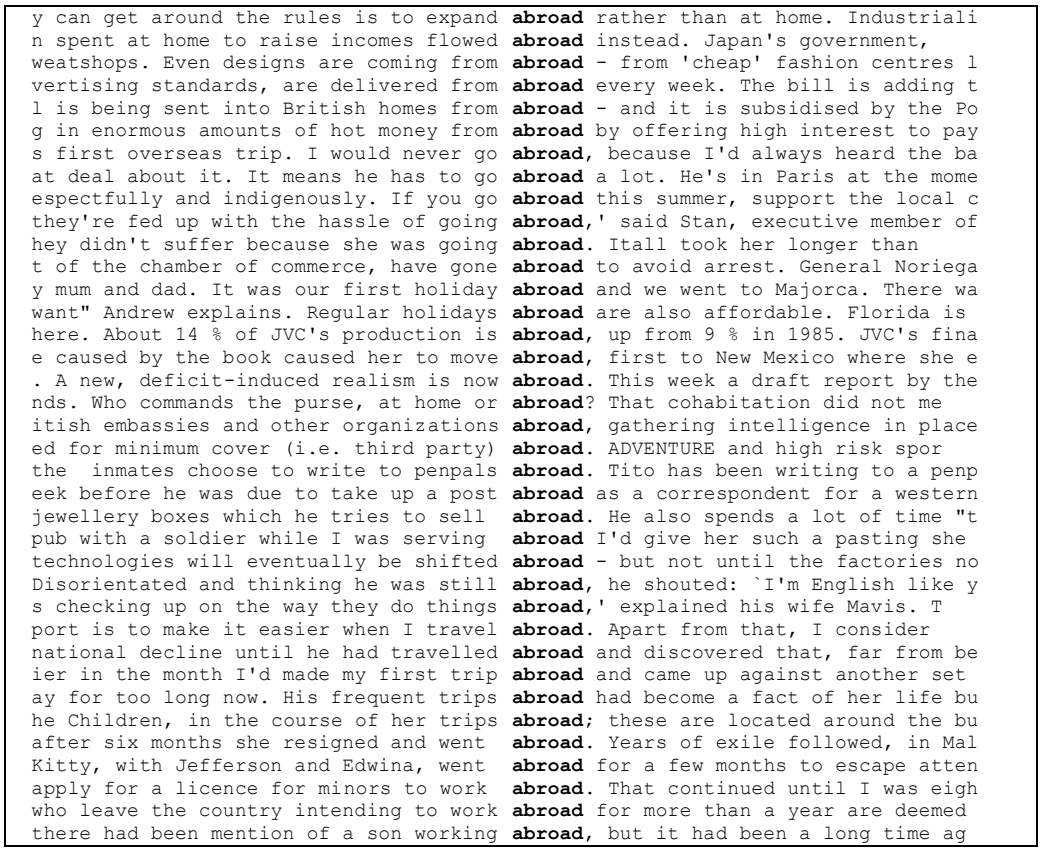
The list gives important information about the basic vocabulary use which can be used as a main source for comparisons. The word 'the', for example, has been used 439,723 times, which is 0.44% of all words in the corpus. By looking at the percentage of our own use of 'the', we can then say whether we have used the word appropriately, naturally, or excessively. If we are comparing or checking, for example, how we use the words in our own essay, a more precise list is needed, that is, the corpus of written language. This is because there is a difference between written and spoken language, including the difference in frequency of 'natural' use of a certain word.

3. Word Concordance for Foreign Language Learning

Many studies concerning the use of concordances and corpus analysis have been conducted in the area of foreign language learning. Tribble in Flowerdew (2002) used a large corpus for teaching academic writing. He used the frequency list of his corpus to help his students learn from native speakers how to use a particular word in sentences. His students also learned a lot about word collocation and how native speakers have used the words in the context. By doing this, a learner will be able to learn to use the target language as native speakers do. Granger's study (2002) also inspires other researchers to use the corpus analysis for instructional aims: teaching a foreign language and even developing a language teaching curriculum.

One important source for learning from a concordance engine is the concordance lines (O'Keeffe, et al., 2007:8). Concordance lines are usually scanned vertically at first glance, that is, looked at up or down the central pattern, along the line of the node word or phrase.The concordance lines provide an example of how a certain word is used, taken from all sentence entries in the corpus. This gives an illustration which enables the users to analyze, copy the pattern, or discuss the use of the selected word. The visual representation of the data makes it easier for the user to analyze and find empirical data concerning questions such as what prepositions usually occur before the word, what are the most frequently used

prepositions, what verb usually proceeds the word, whether the word can end a sentence, and the like. The concordance list in Figure 2 shows how the word 'abroad' has been used by the language users, and therefore tells us how we should use it in our own production.



y can get around the rules is to expand	**abroad**	rather than at home. Industriali
n spent at home to raise incomes flowed	**abroad**	instead. Japan's government,
weatshops. Even designs are coming from	**abroad**	- from 'cheap' fashion centres l
vertising standards, are delivered from	**abroad**	every week. The bill is adding t
l is being sent into British homes from	**abroad**	- and it is subsidised by the Po
g in enormous amounts of hot money from	**abroad**	by offering high interest to pay
s first overseas trip. I would never go	**abroad**,	because I'd always heard the ba
at deal about it. It means he has to go	**abroad**	a lot. He's in Paris at the mome
espectfully and indigenously. If you go	**abroad**	this summer, support the local c
they're fed up with the hassle of going	**abroad**,'	said Stan, executive member of
hey didn't suffer because she was going	**abroad**.	Itall took her longer than
t of the chamber of commerce, have gone	**abroad**	to avoid arrest. General Noriega
y mum and dad. It was our first holiday	**abroad**	and we went to Majorca. There wa
want" Andrew explains. Regular holidays	**abroad**	are also affordable. Florida is
here. About 14 % of JVC's production is	**abroad**,	up from 9 % in 1985. JVC's fina
e caused by the book caused her to move	**abroad**,	first to New Mexico where she e
. A new, deficit-induced realism is now	**abroad**.	This week a draft report by the
nds. Who commands the purse, at home or	**abroad**?	That cohabitation did not me
itish embassies and other organizations	**abroad**,	gathering intelligence in place
ed for minimum cover (i.e. third party)	**abroad**.	ADVENTURE and high risk spor
the inmates choose to write to penpals	**abroad**.	Tito has been writing to a penp
eek before he was due to take up a post	**abroad**	as a correspondent for a western
jewellery boxes which he tries to sell	**abroad**.	He also spends a lot of time "t
pub with a soldier while I was serving	**abroad**	I'd give her such a pasting she
technologies will eventually be shifted	**abroad**	- but not until the factories no
Disorientated and thinking he was still	**abroad**,	he shouted: `I'm English like y
s checking up on the way they do things	**abroad**,'	explained his wife Mavis. T
port is to make it easier when I travel	**abroad**.	Apart from that, I consider
national decline until he had travelled	**abroad**	and discovered that, far from be
ier in the month I'd made my first trip	**abroad**	and came up against another set
ay for too long now. His frequent trips	**abroad**	had become a fact of her life bu
he Children, in the course of her trips	**abroad**;	these are located around the bu
after six months she resigned and went	**abroad**.	Years of exile followed, in Mal
Kitty, with Jefferson and Edwina, went	**abroad**	for a few months to escape atten
apply for a licence for minors to work	**abroad**.	That continued until I was eigh
who leave the country intending to work	**abroad**	for more than a year are deemed
there had been mention of a son working	**abroad**,	but it had been a long time ag

O'Keeffe, et al. (2007:16)

Figure 2. Concordance list of the word 'abroad'

The advantages that a concordance gives are obvious. Because language learning needs exposure to the real use of the target language, a precise source will lead the language learner to the correct use of the language. Below are some advantages that we can take from a concordance engine result.

a. It provides interesting and authentic examples versus traditional examples.

The information about how the language users have used the words in their production is authentic and provides a useful insight for learners. They can compare their own use to the data in the concordance. If the proportion of the use of a certain word is too far from that in the concordance, we can simply conclude that our production is also far from the state of being natural or native-like. Traditional examples of language use drawn from intuition or self-experience might lack the naturalness of the language use. The word 'kill',

for example, surprisingly belongs to the list of rarely-used words, and therefore becomes not too important to be introduced to beginner users of English. The conjunction 'and', in contrast, is the most frequently-used conjunction, far leaving other conjunctions in the list. Therefore, it should be first introduced or taught.

b. It can be used to self-check our own production.

The natural use of language is one of the characteristics of intermediate of advanced language users. By comparing our production to the data in the concordance, we can judge whether we have used the language in a natural manner. Checking our composition or other pieces of writing, i.e., analyzing it in a concordancer and then comparing the result to the data resulting from a large corpus, is important; it can tell whether we have produced a piece of writing that is similar to or different from the writings of many other people's. If the words that we use have very different percentage of occurrence compared to the concordance, we can be sure that we are still far from being a perfect, natural language user.

c. It can be used to check the naturalness/resemblance of our language use if compared to how native speakers have used the language.

Checking the resemblance of our language use to the native speakers' use is one way to learn the language. The resemblance can be seen from many aspects that a concordance engine can tell us, such as from the occurrence of certain words, the similar use of certain phrases, the collocations, or the choice of words. Collocation can best be learnt when we compare how we use it to how native speakers do. Because the real use of language becomes the main source of comparison, a high degree of similarity will mean a lot concerning our efforts to become 'natural' users.

d. It can be used to study the grammar pattern of a language (find the rules, verb patterns, noun phrase orders, and the like).

Because grammar is like the skeleton, a weird pattern in our production suggests that we are not good users and our production is less accurate and, to a certain extent, less intelligible. The inappropriate use of a certain pattern can be minimized by looking at the sample provided by the concordancer. Observing how the word 'will' is used in the concordance data, for example, can lead to proper use of the word in our own sentence. A grammar lesson in which the examples are taken from the real use becomes a powerful source of information concerning the language rules and patterns.

e. It can be used to select words to be introduced to a certain group of learners in a vocabulary class.

When a teacher must select which words will be introduced for a beginner group of language learners, sometimes s/he cannot decide easily. The choice should be based on the real need of the learners for communicating their ideas, and often the needed words are not among the list that the teacher makes. The concordancer provides a precise list of the necessary words, because the list is developed based on the data of the real use of the language. When the list is based on the teacher's own experience or other questionable sources, there is a possibility that the learners are exposed to rarely-used expressions that they themselves may never use them for communication.

f. It helps teachers write appropriate books or materials for a particular segment.

When teachers are going to develop their own materials for their students' learning, the concordance supplies them with suitable data. Modifying an authentic material for teaching beginners is not an easy task. Simplifying a complicated text containing a number of difficult words is unavoidable when we want to use the text for teaching lower level learners. For teachers, this job can be frustrating, especially when they are not assisted with a handy list containing simple words. This list can actually be taken from the concordance engine. Because the list contains only most important words, taking the words from the list to replace difficult vocabulary will make the text simpler, and therefore it becomes suitable for lower level learners. Getting a list of words from another source does not always help, especially when the source was not originally based on the real language use.

4. Towards a Large Indonesian Corpus: A Big Challenge for Indonesian Linguists and Educators

The big challenge to Indonesian teachers, writers, and linguists lies on the unavailability of a large corpus of Indonesian texts. The effect is somewhat huge: textbook writers are stuck to the excessive use of intuition and feelings when selecting sample words or the wordlist to be learnt by students. Bilingual dictionary writers cannot have an easy access to the data about the most frequently-used words or most needed words for their Indonesian-English dictionary entries. In order to write a really relevant and handy, short pocket bilingual dictionary, the writer must search for the most needed expressions, words, or entries. The limited space must not be wasted by including rarely used words in the list. The

intuitive decision to cross out some words from and to take some other words into the list will result in irrelevant entries in the dictionary. There is a big possibility that some needed, high frequency words are missing in such a dictionary, replaced by some unneeded words printed among the entries.

The above illustration will never appear if a large corpus of Indonesian texts is available. Some universities and language researchers and linguists may have begun this big project. However, to produce a big, highly reputable corpus, they still have a long way to go. Changing a large number of written texts into digital images is not an easy job. Collaborations among many different parties might be needed, and financial supports from the government will help.

A more difficult task will be in the area of spoken corpora. The data of spoken language are very rarely found in written forms. Providing the database of this spoken language mode therefore will take a longer route. Collaborative work between associations of researchers, universities, and other parties is needed. Probably there have been several small corpora developed individually by several different institutions, and a joint project to make a large, national-level corpus will provide us a much better data source.

5. Conclusion

The use of concordance engine for language learning is not a new issue, especially in English speaking countries. In such countries, large corpora of different varieties of language use have been developed for different purposes, mainly in linguistics areas and language teaching. However, the trend does not seem to reach most language teachers in Indonesia; discussions on corpus analysis and concordance have been very little.

The advantages that language teachers can get from this program are numerous. Furthermore, it is easy for even a novice computer user to use the word analysis software. The software is available in different forms, either free or paid. The access to a large corpus might be limited, but we can also use our own corpus, even we can use our own writing as the corpus. Analyzing our own production can tell how well we have used the language, and it also enables us to make a comparison between our own production and how English users all over the world have used the language.

The problem for Indonesians lies on the unavailability of large Indonesian corpora, which makes it impossible to analyze, compare, or study the real Indonesian use. It is a challenge for everyone to develop a large, integrated national corpus of the Indonesian

language use. When it is available, the study of Indonesian and foreign language learning for Indonesian students will be much more meaningful.

References

Flowerdew, J. 2002. *Academic discourse*. London: Longman.

Granger, S., et al. 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam: John Benjamins.

Griffin, G. (Ed.) 2005. *Research methods for English studies*. Edinburgh: Edinburgh University Press.

Lamy, M.N. and Mortensen, H.J. 2009. *Using concordance programs in the Modern Foreign Languages classroom* from http://www.ict4lt.org/en/en_mod2-4.htm, accessed 22 Agustus 2009.

McCarthy, M. 2004. *Touchstone: From corpus to course book*. Cambridge: Cambridge University Press.

Mohammadi, M. 2007. "Specialized Monolingual Corpora in Translation", *TJ Interactive: Translation Journal Blog,* Volume 11, No. 2, April 2007.

O'Keeffe, A., et al. 2007. *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press.