**RETRIEVING PATTERNS FROM A CORPUS FOR LANGUAGE LEARNING**

: **ONE STEP CLOSER TO BE AUTONOMOUS LEARNERS**

Prihantoro

Faculty of Humanities, Universitas Diponegoro Semarang

[prihantoro2001@yahoo.com](mailto:prihantoro2001@yahoo.com)

Abstract

These days, teachers are not the only resource where the students can rely on. Existing websites and software are available; hence, it can support the students to be autonomous learners. In the field of corpus linguistics, some corpora are specialized for language learning. Even for general corpora, in the hand of corpus processing software user, they can be a great potential for learning resources. This paper aims at demonstrating how a corpus processing engine and can perform some Natural Language Processing (NLP) tasks that are beneficial for language learning. As for this paper, the corpus and its processing are all performed on-line: the corpus that I use is COCA (Corpus of Contemporary American English. COCA is composed of text from diverse ranges of genres both spoken and written. Students of English department are required to consult this corpus, instead of dictionary, by applying patterns that they find difficult to understand (on the basis of academic writing exercise result).There are some structures that are frequently used by the students and they are encouraged to: 1) test whether the structure patterns are grammatically correct, or ill formed, 2) identify the contexts where the patterns are used, and 3) determine the proper use of the structure by textual evidence. The research procedure is as follows: students are required to input syntax queries to COCA. As it is well understood for COCA users, this corpus processing engine uses regular expression as input queries. These inputs must comply with COCA’s syntax. Three queries are selected to apply, and they are related to the following lemmas <V:suggest>, <V:help> and <N:goal>. Form the concordance display, students can: 1) identify the left and right context, 2) recognize recurrent patterns, 3) associate the context to the text genre, 4) compare with Indonesian corpus, and 5) derive conclusion. In turn, the purpose of this investigation is to prepare them to be both autonomous students and researchers as well.

1. **INTRODUCTION**

The support of computer and internet technology has significantly contributed to the development of all branches of linguistics, including corpus linguistics. Brown corpus in the 1960’s had received a great deal of critics from rationalists, where one of them is Chomsky. He addressed on the processing part that was very laborious as it was performed manually (at that time). See Ludeling & Kyto (2009). However, these days, the criticisms no longer apply as some natural language processing tasks such as part of speech tagging, parsing, and concordance, can be performed easily with the support of computer technology.

In the field of language learning, computer and internet technology support the students to be autonomous learners (Erben, Ban, & Castaneda, 2009). The learners optimize his or her capacity with the least support from human teachers. Teachers are no longer the sole consultant when students encounter problems. In this way, students are practicing self-discovery. When encountering problems, they will identify it and find existing resources that can help them to solve the problems.

In the field of language learning, I have to admit that the rationalists’ method is less complex to apply. The existing template is applied by filling the slot with possible lexicons that comply with the syntagmatic and paradigmatic rule. We often encounter this in most grammar books where examples are made up by the grammarians instead of being taken from language in use. To beginners, this method works best. However, as the learners capability improves, they have to know how language is used in the real world, as it is possible that patterns that they generally apply has some limitations, or structures that are strongly forbidden can be used in the real life. One of the instances is the use of personal pronouns in academic writing as observed by Harwood (2005) and Hyland (2002).

As for this, learners need data that they can work with. As for this, the purpose of this paper is to show how the existing corpus processing engine can help students to perform their own analysis, extracting data from the language in use, and in turn, they themselves can discover how the patterns apply in the real language use.

1. **CORPUS AND CORPUS PROCESSING**

Corpus and corpus processing are two different, but much related, terms. Corpus is the data and corpus processing is how the data in the corpus can be processed. Corpus is linguistic data that is already structurally organized in a way that researchers can process it. However, the processing itself is different issue. In order to process the corpus, a corpus processing engine is required. The engine might be built online or off-line. As for this research, it focuses on one of the online English Corpus, COCA. The engine is already built in the website; so that the corpus is searchable, (some of the corpora are not searchable. COCA or the Corpus of Contemporary American English are available online, created and maintained by Mark Davies, where the tokens are up to 450 billion. This corpus is, up to now, the largest searchable English corpus.

The type of the corpora and the way users use the corpora are quite distinctive. Consider that there are some general corpora such as: British National Corpus, Corpus of Contemporary American English, King Sejong Corpus (for Korean) and Malay Concordance Project (for Malay). These corpora (regardless of the language) are composed of diverse range of texts. Some other corpora are specialized. The collection of this kind of text is specific, such as: Michigan Corpus of Academic Spoken English, Michigan Corpus of Upper Level Student Papers, and Corpus of Historical American English (all of them are English corpus). But, even the general corpus can be used for, one instance, language learning purpose. See Aijmer (2009).

The practice of using corpora for Indonesian students as English learners was demonstrated by Prihantoro (2012). Even though he mentioned that the processing of the corpora can be performed on-line, the focus of his paper is on off-line corpus processing by using a corpus processing software called UNITEX, where the text is collected from a classic english novel. There are also some other works where it focuses on the use of corpus for the purpose of teaching English to the speaker of other languages, such as work by Campoy-Cubillo (2010) which are mostly for Spanish. Gavioli (2005) involved her (Italian) students in corpus investigation for the purpose of language Learning. Therefore, there is still an opportunity to explore how Indonesian students make use of corpus to support English language learning. The author in this case takes the opportunity to show how students use a corpus to explore patterns on their own paces.

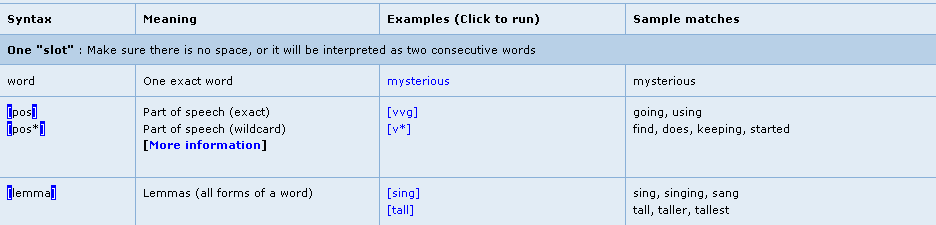
1. **METHODOLOGY: PROCESSING WITH COCA**

We have discussed that the corpus data needs further processing. As for this section, I will show how the data is presented to the users. In the study of corpus linguistics, the result of the processing is presented by concordance. Concordance is the presentation of target query (it can be word, phrases, affixes) where the query is surrounded by the contexts where the query is used in the text corpus. When the corpus, or the corpora, is collected from diverse range of texts, some processing engine display the source texts as well. Concordance is also used in the studies of word frequency, stylistic and discourse as well.

The research procedures in this paper can be described in three layers, following the work of Mc Enery & Hardie (2012), with the application on COCA (Davies, 2011). The first layer is typing input query. The second layer is obtaining concordance and identifying its left-right context. And the third one is finding recurrent patterns and deriving conclusion.

Query is the key to successful retrieval. Therefore, the users have to ensure that the query matches what the users want to look for. The input method of query in COCA is by regular expression, which means the query is composed of the concatenation of character strings (including symbols). The concatenation is often referred as syntax. Regular expression is not the same like the language we use every day. For instance, if we want to retrieve the word forms of <sing>, which are <sing, sings, sang, sung> we have to surround our query with square brackets. Consider figure 1:

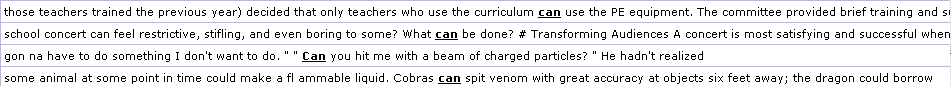
**Figure 1. Samples of COCA Syntax**



The COCA syntax presented by figure 1 describes three descending queries in the following order: retrieve the exact word, retrieve the POS, and retrieve all word forms. It is obvious that the use of the symbols such as brackets and asterisk is not common in the syntax of everyday use. However, this is the syntax that the computer understands (or the syntax that is designed by the programmer). When the syntax is wrong, the computer will reject it and require you to input the right syntax. When the query is composed of proper syntax, it will display the result in the form of concordance.

The aim of the concordance is to show how language is used in contexts. Therefore, the context itself is, sometimes, extremely more important than the query. Concordance displays left and right contexts. Consider figure 2:

**Figure 2. Concordance of <can> as a Verb or a Noun**



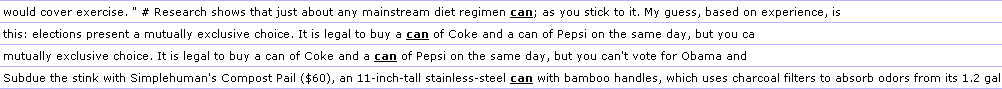


Figure 2 shows the concordance where the target query is <can> (the target query is boldfaced). Consider the first four lines where the right context of the query shows verb (to use, to be, to be, and to hit). This indicates that <can> in this context is a modal auxiliary. The other three lines suggest that <can> is a noun. Consider the left context of the fifth line, where <regimen> functions as attributive modifying <can>. That <can> is a noun is clearly shown by sixth line where it begins with article ‘a’. As it is well understood, in English, the query <can> may have multiple POS, one as a modal auxiliary and another one as a noun. Students can actually find this information when they consult any standard English dictionary. Standard English dictionary these days are corpus based, which means the examples are not made up by the author, but collected from a corpus. When the dictionary authors browse concordance, they identify the left and right context, find recurrent patterns, and derive conclusion. That’s how the corpus based dictionary meaning and examples are created. The next section presents students investigation result of structures related to the following lemmas <V:suggest>, <V:help> and <N:goal>.

1. **COLLOCATION PATTERNS**

**4.1 Suggest that you do or suggest you to do**

The verb pattern for <suggest> seems to be one of the challenges for the students. The results of abstract writing assignment on academic writing course conducted on the first semester of 2012/2013 has shown that 96% of the students use the following lexical chains <suggest><PRO><to><V>. Consider the extracts as shown by example (1) to (3):

1. Therefore, the author suggests the students to do the tasks
2. I suggest them to watch the video first, then ….
3. The program will suggest them to press the red button that will send the stimuli

The result indicated that 100% of the students used <suggest> in the sense of ‘putting forward a plan or idea’, and 96% of the students use this lexical chain. That fact that students only use <suggest> in this sense (whereas the word <suggest> can also be used in another sense), is an interesting phenomena to research, but in this paper I focus on the structure of how they are used. Only two students or 4% of the total students used <suggest> in the lexical chains of <suggest><that><PRO><Vinf>. Please consider examples (4) and (5):

1. However, I have suggested that the students begin from the easiest ones
2. The teachers suggested that the students do classroom observation before ….

Instead of judging the correctness of using one structure over another, I ask the students to browse the corpus to find out 1) the structures in use, 2) the context of how the structures are used. The students begin by inputting the lemma <suggest> plus <that> or <PRO> to the searching box. We will begin by <suggest><PRO><to><V>. The syntax of the query is this = [suggest] [p\*] [to] [v\*]. Consider the result as presented by figure 3:

**Figure 3. Concordance of <suggest><PRO><to><V>**

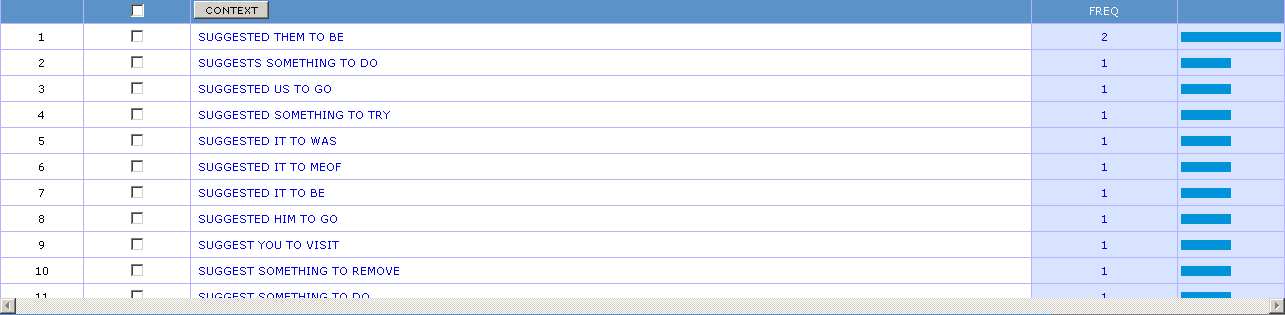
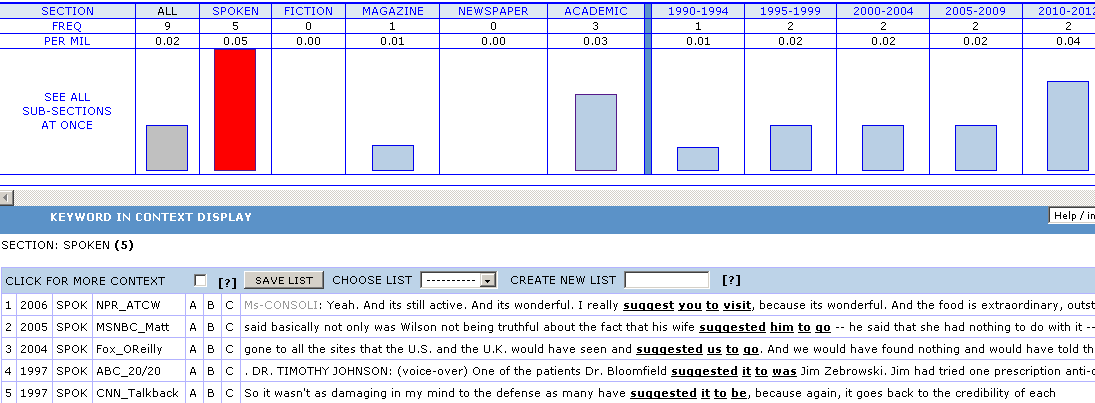


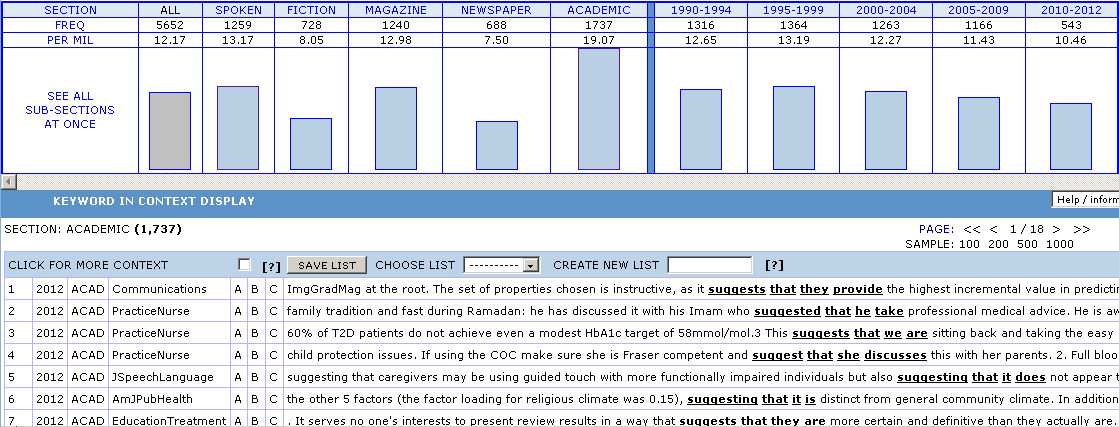
Figure 3 shows that the patterns are in actual use, but very low in frequency. The figure indicates that they are used only one or two times. At this point, I ask the student to refine the search by focusing the pronouns on personal pronouns (you, him, them, us and etc). The syntax is as follows = [suggest] [pp\*] to [v\*]. The result of the retrieval is shown by figure 4:

**Figure 4. Concordance of <suggest><Pers.PRO><to><V>**



The result of the retrieval indicated by figure 4 shows that the pattern is most frequently used in spoken corpus. Now, let us consider how another structure <suggest><that>, is used. The syntax is as follows = [suggest] that [pp\*] [v\*]. Consider the result as presented by figure 5.

**Figure 5. The Source Text of the Concordance <suggest><that><PRO><Vinf>**



By focusing on the source texts, students can derive conclusion. See figure 5, the concordance <suggest><that> are all collected from written academic English, while the concordance <suggest><PRO><to> are all collected from spoken English. As for this, students can derive that the structure <suggest><N><to> is not incorrect. This structure is used, but restricted to spoken English. On the other hand, the structure <suggest><that><N> is widely used in written academic English. Therefore, the judgment is not right or wrong, but more on to which genre each structure applies.

**4.2 Help do or help to do**

The previous sub-section has described how two structures differ under the influence of the text type (spoken and academic). Both structures are correct, but one is more frequently used on the written academic discourse, while another is more on spoken discourse. At this point, students are required to test the other two structures. Consider example (6) and (7) that are obtained from students works (see underlined):

1. The results help the author take conclusion
2. Teachers are not supposed to help the students to do the assignment

Student works indicate the two structures <help><to><Vinf> and <help><Vinf>, are used inconsistently even by individuals. This means that one students use these two structures. Because in the previous sub section, there is a tendency for one structure to be used in different text type, they assume that in this case, one structure also prefers one text type over another. However, we need to test this assumption. The first structure is <help><N><Vinf>. It is expressed by the following syntax = [help] [nn\*] [v\*]. The result is shown by figure 6:

**Figure 6. The Concordance of <help><N><Vinf>**

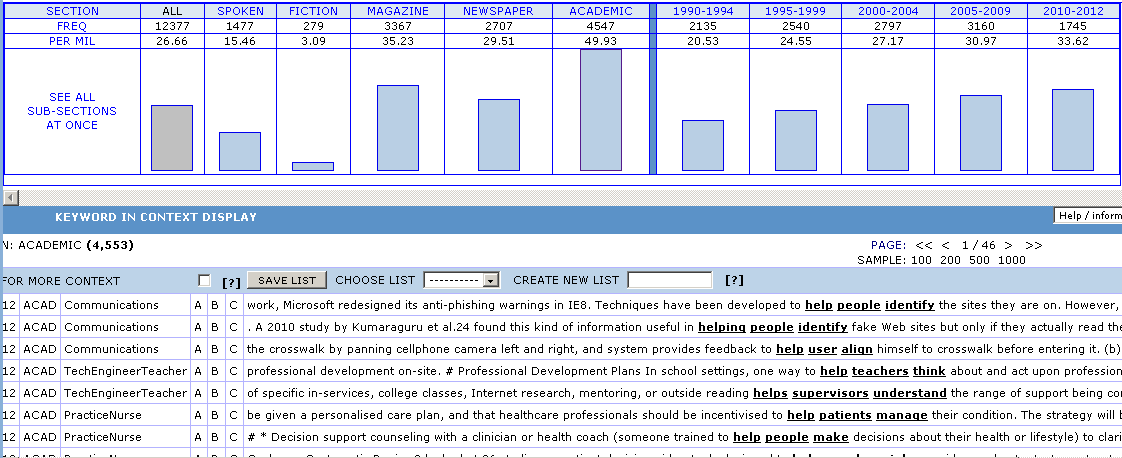


Figure 6 suggests several things. First, historically, there is a steady rise from 1990 to 2012 on how the structure is used. Second, the structures are in actual use. Third, this structure is most frequently used in academic setting. At this point, students are urged not to jump to the conclusion that this is the structure for academic setting while another structure is in the spoken discourse. Students are required to perform the retrieval on another structure, which is <help><N><to><Vinf>. This structure can be expressed by the following syntax = [help] [nn\*] to [v\*].

Students are required to observe the possible structure involving the lemma <help>. Lemma <help> is entered into the query box. The computer will perform automatic retrieval to the lemma, and the result will be shown in the form of concordance. The focus of the retrieval is shown by boldfaced words, but what makes it useful is the left and right context of the target word/s. As for this research, the students are requested to focus on the right context. See figure 7:

**Figure 7. The Concordance for <help><N><to><Vinf>**

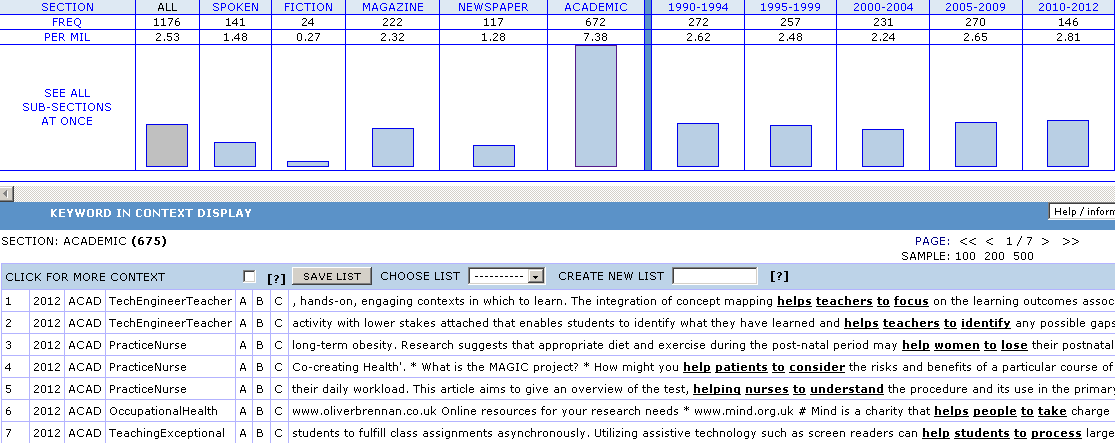
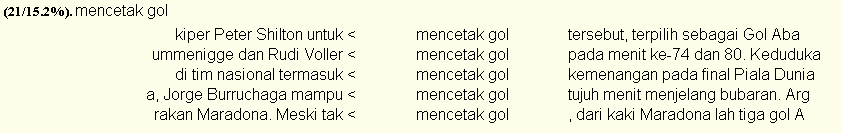


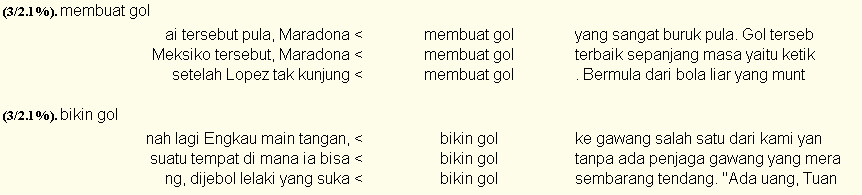
Figure 7 suggests several facts. First, the structure is in actual language use. Therefore, none of the two structures is ill-formed. Second, the structure is also used in the academic setting. Therefore, it rebuts the assumption that one structure preferred on one genre over another. Both structures can be used in academic setting. Third, unlike <help><N><Vinf> that undergoes a steady rise since 1990, the use of <help><N><to><Vinf> seems to be consistently used over the past 22 years. This indicates that the previous structure is gaining popularity over this structure. Even though the structure (<help><N><to><Vinf>) is also used in academic setting, but the trend these days is to use (<help><N><Vinf>). This suggests that besides the frequency of recurrent patterns does not only reveal the text type preference, but also trend over time.

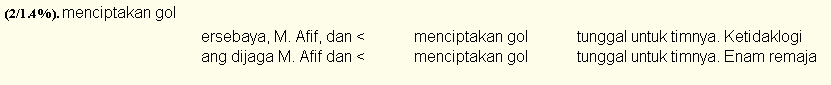
**4.3 Make goal, score goal, or print goal?**

The noun <goal> may refer to points scored by the team player in a sport game. The most frequent recurrent pattern is <NP><V><goal>. The pattern <V><goal> is interesting to research as students have the tendency to choose verbs on the preference of Indonesian collocation patterns. The results of pattern matching of <V><*gol*> from Indonesian corpus indicated the same result. There are four verbs in Indonesian that collocates with <*gol*>, which are: <*mencetak*>, <*membuat*>, <*bikin*> and <*menciptakan*>: they can be translated literally as: ‘to print’, ‘to make’, ‘to make (inf), and ‘to create’. Consider figure 8 that is obtained from Sealang Indonesian Corpus:

**Figure 8. <V><gol> pattern in Indonesian Corpus**



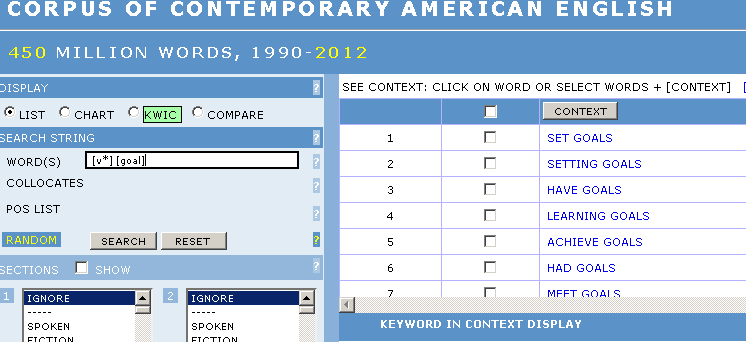




The concordance result has shown that there are four verbs that collocate with <gol> in Indonesian. These verbs are often preferred by the students, but the strings resulted by the collocation patterns are odd as they use the literal translation of those verbs.

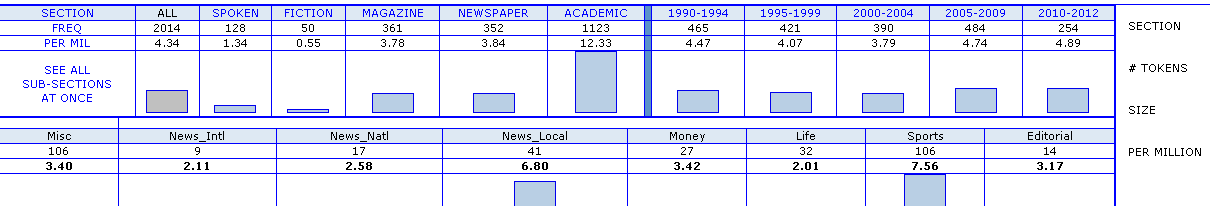
In order to understand the kind of verb used to collocate with <goal>, pattern <V><goal> is applied on COCA. Square brackets are used for each token to indicate that all word forms are retrieved, which will include all verbs (present, past, infinitive) and all word forms of goal <goal, goals>. The result of the retrieval, as shown by figure 9, indicates that <goal> has different senses. It does not necessarily relate to <goal> where points are earned. Therefore, the verbs also vary. See figure 9:

**Figure 9. Concordance < V><goal>**



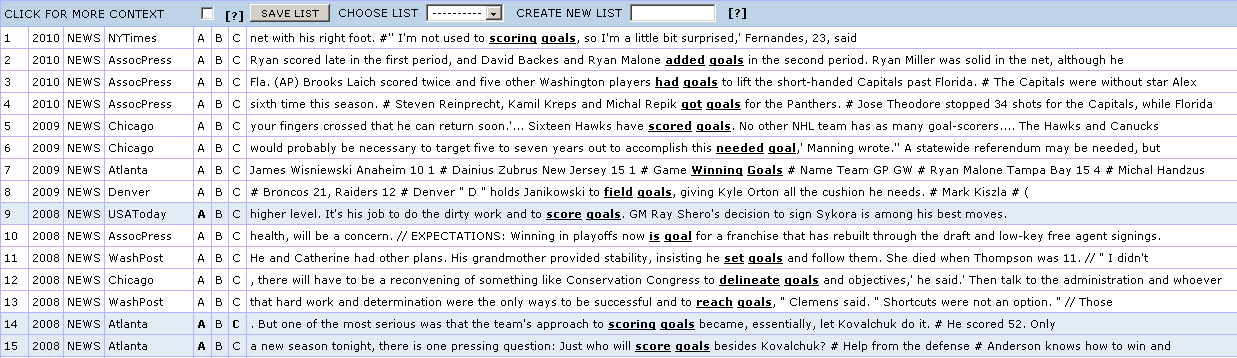
The top six verbs do not seem to be the verbs that suffice the sport-definition of goal, which refers to points scored by the team players. They are more likely to be the representation of <goal> as a purpose. COCA allows the user to view the distributional chart of the pattern. If you check on figure 10, you can understand that <goal> is frequently used in the genre of ‘academic’ as compared to others.

**Figure 10. The Distribution Chart of <V\*><goal>**

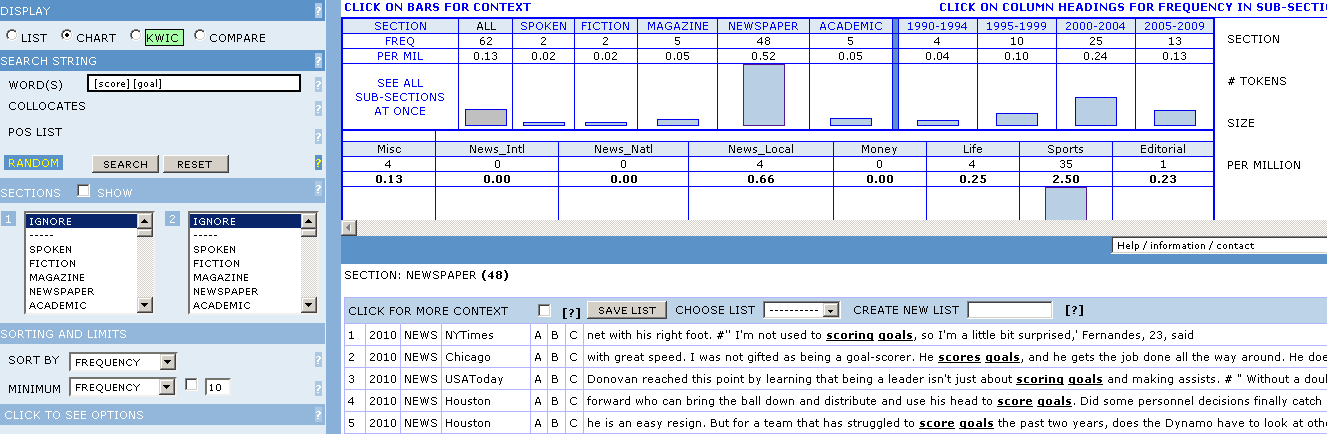


To make sure that the research purpose is achieved, we focus our search on news paper corpus, in sport section (because that is what we look for). The sport section as it has been predicted is the corpus where the pattern frequently occurs. Figure 11 shows the result of the refined retrieval which specifically aims at newspaper corpus. You can see on the left side that the newspapers vary from the sources (NYTimes, Association Press, USA today, Washington Post and etc), or from the section (national, international, local, money/finance, life, sport and editorial). Consider figure 11 where <goal> is used in magazine corpus, and more specifically on sport section (figure 12):

**Figure 11. <V\*><goal> in Magazine Corpus**



**Figure 12. <V><goal> in Sport Section of Magazine Corpus**



This refined search seems to generate positive result. As we specify more and more, we begin to understand that the pattern <score><goal> is the most frequent (and the only) pattern (which refers to the sense we are looking for) and it is most likely to appear in sport section. The corpus indicates that <score> is the only verb to collocate with <goal>. This verifies the assumption that the preference of other verbs such as <creates, make> is the influence of students’ first language, in this case Indonesian.

1. **CONCLUSION**

Computer and internet technology significantly support students to be autonomous learners. In the field of language learning, students can make use of existing corpora to investigate patterns and compare with grammar textbooks or dictionaries where some of the examples are made up by the author instead of corpus based. The dynamics of the corpus such as the text type, time frame, and frequency provide a considerable support for the students to perform investigative exploration. By observing the concordance, with proper input query, they are expected to identify left and right context, find recurrent patterns and be able to derive conclusion based on the corpus evidence.

# REFERENCES

Aijmer, K. (2009). *Corpora and Language Teaching.* Amsterdam: John Benjamin Publishing.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Campoy-Cubillo, M. C. (2010). *Corpus Based Approaches to English Langauge Teaching.* London and New York: Continuum International Publishing.

Charles, F.-M. (2004). *English Corpus Linguistics.* Cambridge: Cambridge University Press.

Davies, M. (2011). The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing* *, 25*, 447-465.

Erben, T., Ban, R., & Castaneda, M. (2009). *Teaching English Language Learners Through Technology.* New York: Routledge.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning.* Amsterdam: John Benjamin Publishing.

Harwood, I. (2005). Inclusive and Exclusive Pronouns in Academic Writing. *Applied Linguistics* *, 26* (3), 343-375.

Hyland, K. (2002). Authority and Invisibility: Authorial identity in academic writing. *Journal of Pragmatics* *, 34*, 1091-1112.

Ludeling, A., & Kyto, M. (2009). *Corpus Linguistics: An International Handbook.* Berlin and New York: Walter de Gruyter.

Mc Enery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics.* Edinburgh: Edinburgh University Press .

Nesselhauf, N. (2005). *Collocation in a Learner Corpus.* Amsterdam: John Benjamins Publishing.

Prihantoro. (2012). On the Use and Design of Corpus for Language Teaching. *Proceeding The 59th TEFLIN International Conference (ISBN: 9786029683943).* Surabaya: Widya Mandala catholic University Press.