THE INFLUENCE OF STUDENTS' L1 AND SPOKEN ENGLISH IN ENGLISH WRITING: A CORPUS-BASED RESEARCH

Prihantoro
(prihantoro2001@yahoo.com)

Universitas Diponegoro
Jl. Prof. Soedharto Tembalang, Semarang, Jawa Tengah

Abstract: Academic writing requires both style and grammatical correctness; however, efforts in improving the quality of English academic writing by non-native students have been focused on grammar. Structures observed in this study were grammatically correct, but considered unnatural in academic writing genre. This research involves a group of non-native English speaking students who were assigned to submit two different kinds of writing to an online repository: a research paper abstract and a free writing article. A survey to understand the sources of English exposure is also conducted. The objectives of this study are to describe unnatural sequences/Multi Words Units (MWUs) used by the students and to identify the motives of using such sequences. The tools for corpus processing used are *Unitex* and *Antconc*. Corpus of Contemporary American English and British National Corpus are also used as reference corpora for English while the SEAlang Indonesian Corpus is used to validate the influence of first language (L1). The analysis of these sequences with comparison to reference corpora indicated the influence of spoken English and students' L1 (Indonesian). This corresponds to the results of the survey that most of the students are exposed to English mostly via spoken, and non-academic sources (songs, movies, social media, etc).

Keywords: corpus, recurrent patterns, lexical bundle, L1, L2, Academic Writing

DOI: http://dx.doi.org/10.15639/teflinjournal.v27i1/217-245

Students pursuing higher education are always required to write academic works such as essays or papers. This can be a challenging task, given that writ-

ten language needs to be distinctive from spoken language (see Biber (2006a); Biber (2006b); and Swales (1990)). For that reason, even a native speaker of a language may find academic writing demanding. As for writing academically in a foreign language, the problem is definitely more complicated. Consider some evaluative expressions that are commonly present in everyday spoken English, but rarely used in academic texts as shown by example (1):

(1) *It's fantastic. That is a fabulous work. I love this method*

Authors of academic writing are most likely to resort to more objective evaluation structures such as: *This research is of a crucial importance. The work is significant,* or *The method is commonly preferred*. As for students whose L1 is not English, the evaluation of the writing is mostly on grammar; which is how the students can express their ideas logically through grammatically correct sentences. However, errors still happen, particularly for beginner students. One of the reasons for these errors to take place is L1 influence (interference or negative transfer) as evidentially shown by Sawalmeh (2013), Isaac (2008), Figueredo (2006) and Arabski (2006). In countries where English is not widely used, the curriculum of English departments normally begin with some basic competence skills, including writing/composition. This writing/composition classes are usually the primary requirements for academic writing. At this point, teachers start to introduce the stylistics of academic writing papers.

The general purpose of this research is to collect data on the academic writing performance of non-native speakers of English, whose L1 is Indonesian. Specifically, it seeks confirmation whether unnatural sequences used by the students are under the influence of L1 or spoken English (or both). The processing is completely carried out by corpus processing softwares and the analysis of this data is cross referenced with two reference corpora to confirm the validity of the author's evaluation. This research also attempts to validate whether performance, as expressed in the writing, corresponds straight to the sources that the students have been exposed to.

Some following sub-sections here also review related studies about interference specifically related to Indonesian as L1. The relevancies of these studies to the present research and in what respect it differs will be explained here. Methodology section describes in detail how this study was carried out. It also describes the tools and corpora used in this study. The core of this study is

presented in the finding and discussion section. While the finding section focuses more on the quantitative analysis, the discussion section explores the data with more qualitative approach. The conclusion section correlates the corresponding qualitative to quantitative natures of the data as well as proposes recommendation for further studies.

Written and Spoken Language: Corpus Validation

The distinction of spoken and written languages is necessary especially in the field of genre analysis and text production, as described by Halliday (1989) and Hasan (1986). They focused on providing concepts and baselines that characterize the difference between spoken and written language. More recent works, such as Biber (2006), Szmrecsanyi & Hinrichs (2008), applied these concepts to more controlled variables, which are college students. The results of research in spoken and written language is also documented in standard dictionares. The documentation of written language in specific domain is usually shown by <written> special annotation to the related entries. Consider the annotation for entry <exclaim>, as shown by figure 1.

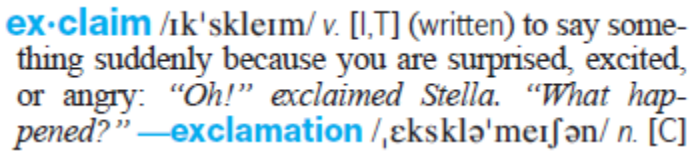


Figure 1: Annotation for Entry <exclaim>

Figure 1 is taken from Longman Dictionary of Contemporary English, where the entry <exclaim> is marked (written) as it is used more frequently in written language as compared to spoken language. However, since features of language may change from time to time, we need to make sure that the dictionary is recent or we can validate this using dynamic corpora (corpora that are regularly updated). Focus on the retrieval of <exclaim> in Corpus of Contemporary American English (COCA) and its genre distinction are shown in Figure 2.

The result indicates that <exclaim> is used widely in fiction and other text types (note that in COCA, besides spoken section, all data are taken from written texts). The lowest frequency is on spoken section. Thus, it validates the claim that <exclaim> is widely used in written language.

Besides the choice of word, the distinction can also be shown by sequence or pattern preference. The term sequence or pattern refers to a string of words, which is considered pre-fabricated expressions. These expressions are syntactically composed, but used paradigmatically. A variety of terms have been proposed. Hyland (2008) and Allen (2009) refer to the sequence as lexical bundle, as also used in Allen (2009) and Chen and Baker (2010). While in the field of computational linguistics, the term lexical chain or multiword units (MWUs) is also used (see Paumier, 2008).

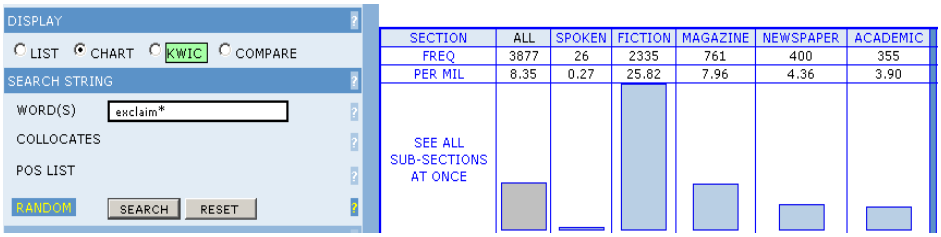


Figure 2: Lowest Frequency of <exclaim> in Spoken Section of Corpus Data

It is important to understand what makes MWUs essential in the field of language learning. Rather than concatenating words by words, using pre-fabricated sequence will reduce the risk of making mistakes. In turn, this will improve the quality of students' academic writing. Besides, this will also help characterize the texts to a specific domain, as the identification of a text to a specific genre may derive from lexical bundles (Hyland, 2008). Meanwhile, Figure 3 below introduces us to the expression <it is commonly believed> which according to COCA has zero occurrences in spoken section, but the frequency is very high in academic section.

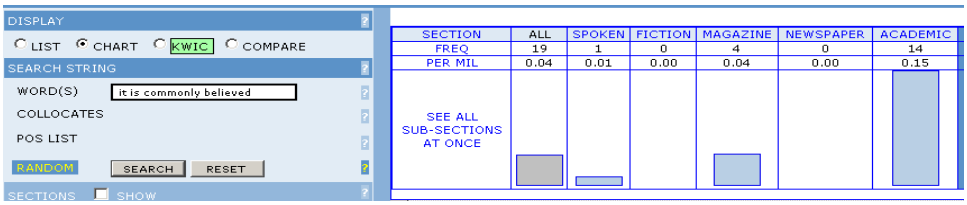


Figure 3: Frequency Chart of <it is commonly believed> in COCA

How corpus data can be used to validate judgment is relevant to my research as the dynamic corpus is always updated to record the actual language use. Besides amplifying the analysis, corpus data is also useful to show the influence of L1. There are some patterns observed in L2 writing that correspond to L1 structure. This early hypothesis can be validated with help of reference corpora.

Studies on Interference

Some of the research in academic writing, especially where the subjects of research are students, whose L1 is not English, is often focused on error analysis. The research usually describes and categorizes students' errors. The categories where errors take place are prioritized target of improvement for teaching. Error analysis patterns may further be compared to the linguistic features of students' L1 to validate whether the error is transferred from L1. When writing in English, for example, it is quite common for Indonesian students to drop articles since this linguistic feature is not present in Indonesian (Wijaya, 2012). In the languages where such markers are absent, similar problems tend to occur (Bautista & Gonzalez, 2008).

Some other academic research focuses not on grammar, but on the style and word choice of the writing. It tries to measure how appropriate the style and the word choice of the writing to the genre where the language is used. One of such research studies the preference of pronoun 'I' that is getting more and more common these days in academic writing (Harwood, 2005).

Some of the works dedicated to the interference of Indonesian in English were conducted by Fauziati (2003), Roni (2006), Yembise (2011) and Moehkardi (2012). While Fauziati's (2003) respondents were middle schoolers, my respondents in this research are all college students. Roni's (2006) respondents were college-level students, but they were not from English majors. In my research, the students were all from English department and they were all senior undergraduate students who received minimum B- (grade) on Academic Writing. Moehkardi (2012) describes some lexical bundles in English and lists some possible L1 transfers patterns. Unlike Mohkardi (2012) whose research did not take account of authentic data, my research is fully data driven, and all the data are processed by using a corpus processing software. None of the previous research employed any corpus processing software and none of them used reference corpus as well. The validation with corpora

reference, not to disregard the introspective competence of the researchers, is of a crucial importance as the variables of corpora are overt, and most importantly, the data in the corpora are authentic, updated and evidential. Reasons for interference to occur may vary, but one classical factor, as also mentioned in the previous studies, is the L1 of the students (Husein & Mohammad, 2012).

Another factor that may contribute to interference is the spoken language. Šimčikaitė (2012) discovers that students use some English spoken discourse markers in writing like *I mean, anyway,* and *by the way*. The interference of spoken language is well documented by Krauthamer (1999). He refers to the interference as SLIP (Spoken Language Interference Patterns). The present study is aimed at testing whether these two factors contribute to students' writing. Cross reference is conducted by using COCA (Davies, 2008), BNC (Aston & Burnard, 1998) and SEAlang Indonesian Corpus (Scannel, 2010).

This research is also digital data driven, which means that all data are digitally processed. The corpus methodology in this research follows the works of Pang (2010), Allen (2009), Hyland (2008) and Yoon (2008), where the recurrent lexico-grammatical patterns were retrieved by corpus processing software and analyzed both quantitatively and qualitatively.

METHOD

The respondents of this study are a group of Undergraduate English Department students of the 6th semester from *Universitas Diponegoro*, Indonesia. I accessed the on-line academic information system and screened the students with the following variables: 1) students who have passed academic writing class, 2) students with at least a final grade of B-. Of around 120 students, 117 passed the first screening, and 79 passed the second one. Of this number, 69 students volunteered. Of the 69 students, 65 responses were collected; two students failed to submit the tasks due to failure in establishing a stable Internet connection while the other two dropped the tasks for unknown reasons. There are two types of responses; the first one is response to questionnaires, and the second one is response to writing tasks.

Questionnaires and Writing Tasks

The questionnaire is crucial as it describes the students' exposures to English via different sources. It will be used further in data processing to con-

firm whether the exposure corresponds to their academic writing skills or not. In the survey, students can score the source of exposure to English from the lowest (1) to the highest (5), as is shown in Figure 4 below.

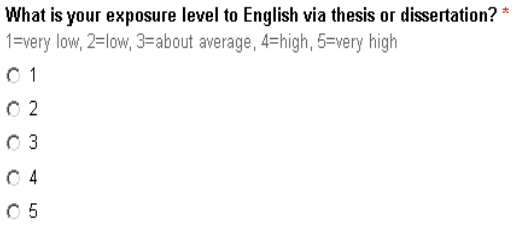


Figure 4: Screenshot of the On-Line Questionnaire

The second crucial task is the task to write two articles. The first one is to write a research paper abstract, and the second one is free writing.

Corpora and Corpus Processing Software

Results of the survey are saved in a spreadsheet file, while the articles are saved in a raw text file allowing *Unitex* and *AntConc*, the corpus processing software used in this research, to further process the texts collection as a corpus. Word Frequency computation and concordance extraction is performed to obtain necessary information that will be presented in the findings section. Analysis to the findings is performed with reference to COCA and SEAlang Indonesian corpus. While *Unitex* is used to analyze the texts used in the research, COCA provides an online reference to standard corpora. The analysis is fully presented in the discussion section. Prior to the writing of this paper, the result of the analysis was presented to the students. Interview session was also held to understand the motivation of why such unnatural sequences were selected by the students. As for COCA, the presentation may include more than concordance lines (see Figure 5).

Figure 5 shows a chart generated by COCA to show frequency of occurrence on the basis of corpus section (left side) and historical trend (right side). The highest frequency corresponds to the other section and it always reaches the top ceiling in the section as shown by academic section on the left side. The historical trend itself is grouped in a four-year term. As a companion to the concordance, this chart takes a very significant role. Actual concordance lines

correspond directly to the frequency. Presenting all concordance lines can be wordy; therefore, I decided to randomly choose any concordance lines to be shown and explained. These selected concordance lines that are presented in the findings and discussion section are representative to the aims of this study.

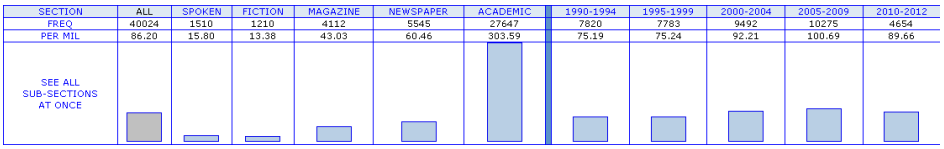


Figure 5: Chart Showing Sections, Historical Progress and Frequency in COCA

FINDINGS AND DISCUSSION

Findings

The survey showed that the students' largest exposure to English included movies, songs and social media (>4). It is important to notice that lecture and assignment are only one level below (3-4). This was expected, as the students are English department students and the classes and assignments are conducted in English. The level of exposure to textbook is 3, which is equal to direct conversations, comic books and newspapers. Theses, dissertations and journals, which are actually significant reference for academic writing are down on level 0-1.

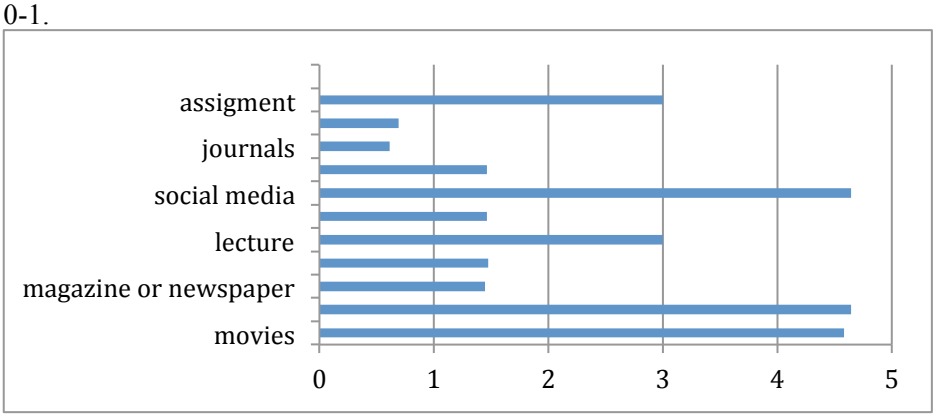


Figure 6: Student Exposure to English

Ironically, students were least exposed to academic sources such as journals, theses and dissertations. Students also claimed that the highest exposure came from social media like *Twitter*, *Facebook*, *Path*, etc. A round score (0) was obtained for corpus as a source of exposure. Overall responses indicated that students were more exposed to non-academic, and spoken languages as compared to academic and written language. This is shown in Table 1 below.

Table 1. Data Segmentation

Data Segmentation			
Written	Spoken	Academic	Non-Academic
31%	69%	37%	63%

The survey also showed that students were exposed more to spoken than written English. I categorized direct conversations, lectures, songs and movies as spoken English, while the rest (journals, theses, assignments, social media, comic books, magazines, assignments and newspapers) as written English. Even though the variables of written English outnumbered spoken English, the average of spoken English (69%) is higher than that of written English (31%).

Token Frequency

Table 2. Token Frequency Extraction

Type	Frequency	Type	Frequency	Type	Frequency
The	481	in	188	English	110
And	219	study	177	Foreign	99
Be	207	language	151	students	97
		but	111	need	90

Token frequency extraction is useful in determining topics and keywords. The result shown in Table 2 suggests a focus on English as Foreign language in the collection of abstracts. It is an interesting notion since the term 'foreign' is used instead of 'second language'.

Lemma Distribution and 3-Gram

While word frequency is useful in determining the topics or keywords, some other means are required to retrieve sequences/MWUs. With this in mind, I decided to first observe lemma distribution. The processing indicates that

there are three lemmas, content words, which occurred in almost each of the students' writing (both abstracts and free-writing) as shown in Table 3.

Table 3. The Distribution of <suggest>, <help>, and <goal>

	Suggest	**Help**	**Goal**
Present	94%	97%	92%
Absent	6%	3%	8%

The identification of sequences used on the left and right context is performed by Local Grammar Graphs (LGGs). LGGs is one of the machine-readable grammars in *Unitex* used to retrieve words or sequences (Paumier, 2008).

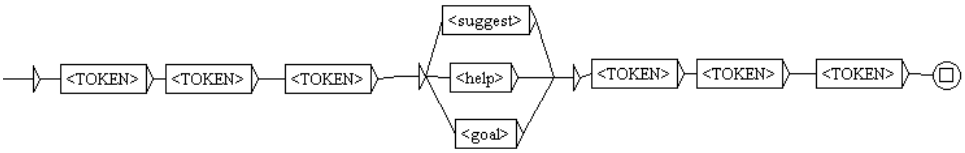


Figure 7: LGGs to Extract 3-Gram Sequences

The LGGs, as shown in Figure 7, extracted three tokens on the left-right context of target lemma <suggest; help; goal>. This process gives reason to name it 3-Gram. When necessary, the retrieval of N-gram (where N may refer to any number) is possible. The angle brackets are required to indicate all word forms of a lemma. As an illustration, the use of angle brackets in <go> retrieve all word forms conceived by corpus-processing software lexical resource which include *go, went, gone, goes* and *going*. LGGs in Figure 7 retrieved 3-gram sequences as shown in Figure 8.

For <suggest>, and <help>, the recurrent sequence patterns are on the right context. The nouns that the verb <suggest> specifies are almost all human nouns, and some are represented by pronouns. After the nouns, <to><Vinf> are used. The same specification also applied for <help>; only, there are two patterns after the nouns, which is to use <to><Vinf>, or just <Vinf>. As for <goal>, it is interesting to observe the left context, especially the verbs that col-

ligate with <goal>. These patterns are explored more with concordance in the discussion section.

Therefore,	the	author	**suggests**	the	students	to
was	finished	I	**suggest**	them	to	watch
The	program	will	**suggest**	them	to	press
as	well	Teachers	**suggest**	this	to	improve
As	many	researchers	**suggests**	the	students	to
not	supposed	to	**help**	the	students	to
now.	This	paper	**helps**	us	to	deal
perform.	The	results	**help**	the	author	take
because	it	will	**help**	them	to	understand
those,	It	will	**help**	the	students	to
very	passionate	to	**goal**	to	equalize	the
amazing!	He	created	**goal**	only	in	his
have	the	same	**goal**	and	equally	work
had	her	private	**goal**.	Without	this	effort
sure.	But	making	**goal**	in	Chelsea	was

Figure 8: The 3-Grams for Left-Right Context <suggest;help;goal> in Abstracts Collection

Students' Feedback

As stated previously, the findings of the study had been presented to the students where they were asked to provide some reasons as to why the suggested structures were not used. Most of the students said that they were not aware of the presence of such structures. As for those who were aware, the reasons for not choosing the structures varied. See Table 4.

Table 4. Awareness of Suggested Structures

	Positive (28%)	Negative (72%)
6%	Strong belief in the source	
10%	Strong belief in frequency	
12%	Strong belief in teacher's instruction	

There are three main reasons why they had the confidence in using the structures, even though they know that the suggested structures exist. The faith derived from consulting, specifically, existing final projects as the source of their writing (6%). The second one is because of the degree of frequency of the overall exposure. The third one is the strong belief in previous teacher's instructions/ descriptions, which are mostly spoken and acquired during the lecture.

Discussion

In this section, I discuss the frequent Multi Word Expressions (MWUs) obtained from the 3-gram extraction of students writing corpus. The MWUs are as follows:

MWU 1	Suggest that you do	- Suggest you to do	
MWU 2	Help you to do	- Help you do	
MWU 3	Score goal	- make goal	- print goal

Should the results of the questionnaires be parallel to the findings, it is possible that domain shift from spoken to written may take place. It is also possible that to some extent, learners replicate the L1 pattern, in this case Indonesian, to English as the foreign language. The description is validated by corpora analysis. In doing so, I consulted two well-known English corpora from two different dialects: British English (BNC) and American English (COCA). To investigate language transfer possibility, the SEAlang Indonesian Corpus is also consulted. Analytical information from each corpus that corresponds to the MWUs will be presented to show the gap between authentic language and language used by learners of English as a second language.

MWU 1: Corpus Data

In this section, I will show how students use the verb <suggest>. The presence of this verb in my corpus is quite significant as students usually use this verb in the end of their writing to give recommendation. The retrieval was aimed at all verb forms of <suggest>: *suggest, suggests, suggested, suggesting*. Among these four forms, 'suggesting' was not found in the retrieval. As for the patterns, there were two verb patterns in use. The first one (---1) was <suggest><PRO><to><V>, and the second one (---2) is <suggest><that><PRO><Vinf>. Figure 9 below shows Concordance 1 that is generated by *Unitex*.

1	Therefore, the author	**suggests**	the students to do the tasks ----------	1
2	was finished. I	**suggest**	them to watch the video first, then`----------	1
3	The program will	**suggest**	them to press the red button ----------	1
4	as well.Teachers	**suggest**	this to improve listening skills, but ----------	1
5	As many researchers	**suggests**	the students to read and listen a lot without	
6	writing styles are	**suggested**	to be analyzed	
7	and students. It	**suggests**	us to believe that ----------	2
8	language, and teacher	**suggests**	that the students do classroom observation ----------	1
9	while the author	**suggests**	the parents to mix codes in the family and ----------	1
10	do. Parents are	**suggested**	to be selective in choosing sources ----------	1
11	However, I have	**suggested**	that the students begin from the easiest ones ---------	2
12	good. The result	**suggests**	the readers to learn phrasal verbs ----------	1

Figure 9: Concordance 1: The Retrieval of <suggest> in Abstracts Collection

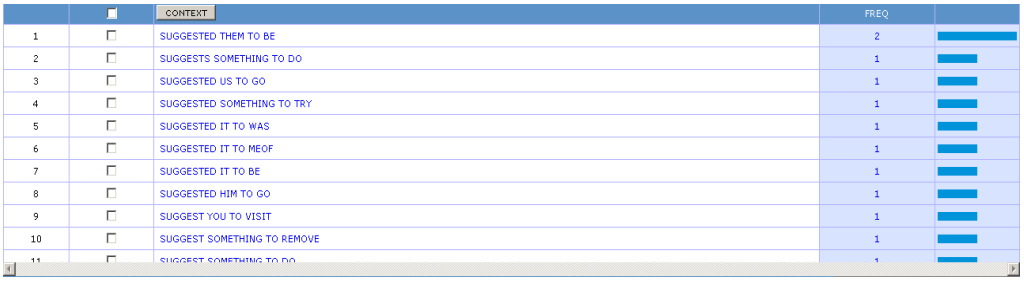
Concordance lines are presented on the left, and pattern variation is shown on the right. Although some other patterns are observed, but these two patterns were quite similar with respect to the nouns that they specify. The nouns that each pattern specifies is human noun, and some of them are replaced by pronoun. Other patterns occur, but the two patterns dominated the use. In the research abstracts, <suggest>Pers.PRO><to><Vinf> is observed to occur more than 75%. The same, even higher, domination was also observed in free writing:

1	the journey. Then he	**suggested**	us to see sunrise in Sikunir ----------	1
2	from. Ancelloti always	**suggests**	Ronaldo to take every free kick ---------	1
3	by local confederation	**suggest**	them the otherwise.	
4	whenever possible. we are	**suggested**	to look for information related	
5	just fine. His boss keeps	**suggesting**	him to work hard and prepare ----------	1
6	in the world must	**suggest**	their children to have ----------	1
7	the coach always	**suggests**	the players to do their best ----------	1
8	the incident, they are	**suggested**	to take leave off the team by the	
9	done. And he also always	**suggested**	me to do houseworks ----------	1
10	matter of time? I cannot	**suggest**	him to take rest, as he is ----------	1

Figure 10: Concordance 2: The Retrieval of <suggest> in Free Writing Collection

MWU 1: Corpus Validations

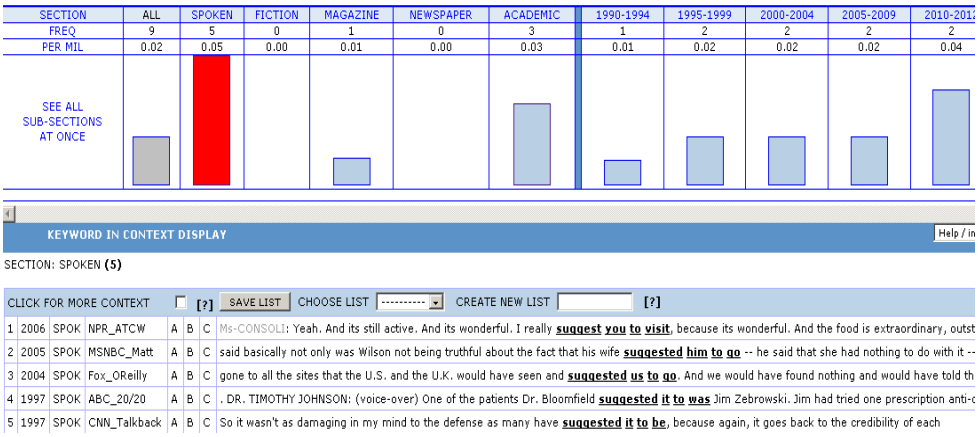
Instead of determining the correctness of particular structures, I navigated the corpus to discover how the structures are present in actual use. The result of the first pattern retrieval in COCA is shown in Figure 11.



		CONTEXT	FREQ	
1	☐	SUGGESTED THEM TO BE	2	
2	☐	SUGGESTS SOMETHING TO DO	1	
3	☐	SUGGESTED US TO GO	1	
4	☐	SUGGESTED SOMETHING TO TRY	1	
5	☐	SUGGESTED IT TO WAS	1	
6	☐	SUGGESTED IT TO MEOF	1	
7	☐	SUGGESTED IT TO BE	1	
8	☐	SUGGESTED HIM TO GO	1	
9	☐	SUGGEST YOU TO VISIT	1	
10	☐	SUGGEST SOMETHING TO REMOVE	1	
11	☐	SUGGEST SOMETHING TO DO	1	

Figure 11: Concordance of <suggest><PRO><to><V>

Figure 11 shows that the first pattern is in actual use, but very low in frequency. The figure indicates that they are used only one or twice. At this point, I then refined the retrieval by focusing the pronouns to personal pronouns (you, him, them, us and etc.):



SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	9	5	0	1	0	3	1	2	2	2	2
PER MIL	0.02	0.05	0.00	0.01	0.00	0.03	0.01	0.02	0.02	0.02	0.04

SEE ALL SUB-SECTIONS AT ONCE

KEYWORD IN CONTEXT DISPLAY Help / in

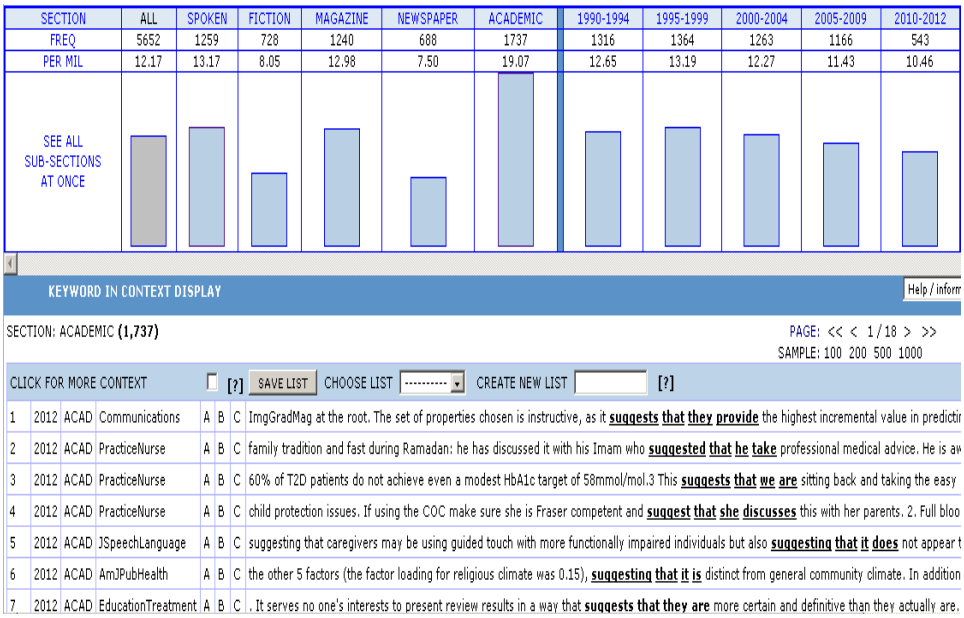
SECTION: SPOKEN **(5)**

CLICK FOR MORE CONTEXT ☐ [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

1	2006	SPOK	NPR_ATCW	A	B	C	Ms-CONSOLI: Yeah. And its still active. And its wonderful. I really **suggest you to visit**, because its wonderful. And the food is extraordinary, outst
2	2005	SPOK	MSNBC_Matt	A	B	C	said basically not only was Wilson not being truthful about the fact that his wife **suggested him to go** -- he said that she had nothing to do with it --
3	2004	SPOK	Fox_OReilly	A	B	C	gone to all the sites that the U.S. and the U.K. would have seen and **suggested us to go**. And we would have found nothing and would have told th
4	1997	SPOK	ABC_20/20	A	B	C	. DR. TIMOTHY JOHNSON: (voice-over) One of the patients Dr. Bloomfield **suggested it to was** Jim Zebrowski. Jim had tried one prescription anti-c
5	1997	SPOK	CNN_Talkback	A	B	C	So it wasn't as damaging in my mind to the defense as many have **suggested it to be**, because again, it goes back to the credibility of each

Figure 12: Concordance of <suggest><Pers.PRO><to><V>

The result of the retrieval indicated by Figure 12 showed that the first pattern is most frequently used in spoken English, and the frequency is very low (5). Further, let us consider how the second pattern <suggest><that>, is used:



SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	5652	1259	728	1240	688	1737	1316	1364	1263	1166	543
PER MIL	12.17	13.17	8.05	12.98	7.50	19.07	12.65	13.19	12.27	11.43	10.46

SEE ALL SUB-SECTIONS AT ONCE

KEYWORD IN CONTEXT DISPLAY

SECTION: ACADEMIC **(1,737)**

PAGE: << < 1/18 > >>

SAMPLE: 100 200 500 1000

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]

1	2012	ACAD	Communications	A	B	C	ImgGradMag at the root. The set of properties chosen is instructive, as it **suggests that they provide** the highest incremental value in predictir
2	2012	ACAD	PracticeNurse	A	B	C	family tradition and fast during Ramadan: he has discussed it with his Imam who **suggested that he take** professional medical advice. He is aw
3	2012	ACAD	PracticeNurse	A	B	C	60% of T2D patients do not achieve even a modest HbA1c target of 58mmol/mol.3 This **suggests that we are** sitting back and taking the easy
4	2012	ACAD	PracticeNurse	A	B	C	child protection issues. If using the COC make sure she is Fraser competent and **suggest that she discusses** this with her parents. 2. Full bloo
5	2012	ACAD	JSpeechLanguage	A	B	C	suggesting that caregivers may be using guided touch with more functionally impaired individuals but also **suggesting that it does** not appear t
6	2012	ACAD	AmJPubHealth	A	B	C	the other 5 factors (the factor loading for religious climate was 0.15), **suggesting that it is** distinct from general community climate. In addition
7	2012	ACAD	EducationTreatment	A	B	C	. It serves no one's interests to present review results in a way that **suggests that they are** more certain and definitive than they actually are.

Figure 13: The Source Text of the Concordance <suggest><that><PRO><Vinf>

The frequency of this sequence is quite high in the corpus. The highest frequency of the concordance <suggest><that> is observed in the academic English section (1737). This indicates that the second pattern is widely used in written academic English, even though both are used. In addition, it is necessary to check the occurrence in another corpus.

British National Corpus (BNC) is another corpus of English that are composed by different text types. In this corpus, the result of the retrieval also showed the same tendency. The pattern <suggest that> displayed 169 hits in academic section, while only 69 hits in spoken section. The retrieval of <suggest><PRO> in academic section resulted in zero hit, while it only had one hit in spoken section and one hit in magazine section.

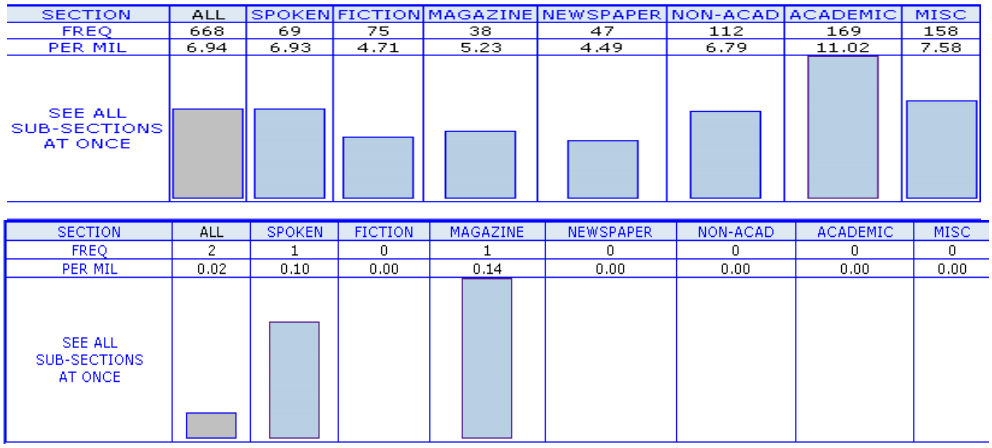


Figure 14: <suggest that> VS <suggest><pro> in BNC

MWU 1: Analysis

Negative transfer may take place not only from first to second language, but also across domains, for instance from spoken to written language. Krauthamer (1999) refers to the interference patterns of spoken language as SLIP (Spoken Language Interference Pattern). It ranges from vocabulary to style. One of the means to help determine whether the writing task is influenced by spoken language is by evaluating its lexical density. For this purpose, I used *AntConc* vocabulary profiler (Anthony, 2006) to measure the density of academic words. The list of academic words (which contains the token and type) is obtained from Coxhead (2000). The list was then improved by Davies (2008) with the help of COCA. The profiling here excludes function words like conjunctions and prepositions.

Table 5. Comparison of Vocabulary Profile: Scientific Writing and Free Writing

Vocabulary Profile (Token of academic word list)	Number	
	Scientific Writing Tasks	Free Writing
More than 60%	4	0
Less than 60%	61	65

Kwary (2013) asserts that when the token percentage of the vocabulary profile is more than 60%, it is statistically significant. Table 5 shows the evaluation of scientific and free writing profile. It shows that the number of scientific writing tasks where the profile receives evaluation more than 60% is only 4, while the number of writing tasks that receives less than 60% is 61. One of the reasons is the use of words that are not included in the academic wordlist such as: *well*, *anyway*, *I mean*. This is in line with Šimčikaitė (2012) when identifying spoken language vocabularies in the writing task. The presence of these words reduces the academic vocabulary densities. Krauthamer's (1999) SLIPs also documented the influence of spoken language stylistics. Nuruzi, Farahani, and Farahani (2012), in studying the stylistics of academic writing, suggests that students use nominalization feature in academic writing.

MWU 2: Corpus Data

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 in the written academic section (another is used more in spoken section).

There are some lemmas in which the word forms are more frequently used in spoken language but less in written language, or vice versa. As for the verb <help>, the structure without 'to' infinitive is less dominant than its counterpart in student's writing corpus.

1	Teachers are not supposed to	**help**	the students to do the assignment ----------	1
2	assessment now. This paper	**helps**	us to deal with linguistic phenomena ---------	1
3	does not perform. The results	**help**	the author take conclusion ----------	2
4	be beneficial because it will	**help**	them to understand more about English ----------	1
5	besides those, It will	**help**	the students to understand conditional sentences----	1
6	plan. The description will	**help**	readers to understand ----------	1
7	foreign language. It will	**help**	them to develop their language performance ----------	1
8	in English. It also	**helps**	them satisfy their individual and social need ---------	2
9	can last forever. It could	**help**	the researcher to get and to find more ----------	1
10	whenever ready. This paper	**helps**	us to know the various codes ----------	1

Figure 15: Concordance 3: The Retrieval of <help> in Abstracts

1	not the goal. It can	**help**	you to increase your leadership skill ----------	1
2	same match. It can	**help**	many students to enrich their knowledges -------	1
3	scores well. They	**helped**	me to believe that I will be a success ------- 1	
4	moment. It will	**help**	you to explore your business potential ------- 1	
5	level up. It can	**help**	us to get a good job---------- 1	
6	fair. The supporters	**helped**	the team to win the match ----------	1
7	now. He cannot	**help**	Madrid to win more trophy ----------	1
8	Then he said, 'god,	**help**	me to survive this 90 minutes ---------- 1	
9	make three goals.	**Helping**	Chelsea to maintain Drogba is useless. ------- 1	
10	position now. I	**helped**	the players to win, and they helped me to stay ------	1

Figure 16: Concordance 4: The Retrieval of <help> in Magazine/News Article

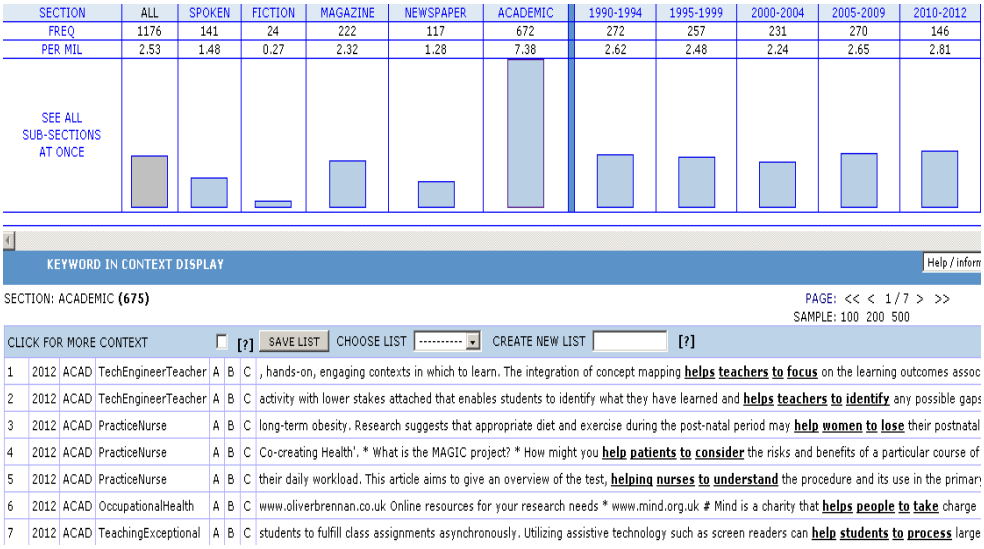
Both concordances indicate two structures <help><to><Vinf> and <help><N><Vinf>. The first one is used more frequently in both abstracts and free writing article. But the second pattern is used in abstracts writing only. Whether there is a preference of one genre over the other still requires a corpus validation.

MWU 2: Corpus Validation

Moving on from MWU 1, this section presents BNC and COCA validation. Consider Figure 17, which presents the concordance of <help><N><Vinf> by COCA:

Figure 17: The Concordance of <help><N><Vinf>

Figure 17 suggests that first, there is a steady rise from 1990 to 2012 on how the structure <help><N><Vinf> is used. Second, both structures are present in actual use. Third, this structure is most frequently used in academic setting (see the highest frequency shown in the academic section (4547)). Further, a validation is required to observe whether the same trend applies to the second pattern <help><N><to><Vinf>. The result of COCA retrieval of this pattern is shown in the following figure:



SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
FREQ	1176	141	24	222	117	672	272	257	231	270	146
PER MIL	2.53	1.48	0.27	2.32	1.28	7.38	2.62	2.48	2.24	2.65	2.81

SEE ALL SUB-SECTIONS AT ONCE

KEYWORD IN CONTEXT DISPLAY

SECTION: ACADEMIC **(675)**

PAGE: << < 1 / 7 > >>
SAMPLE: 100 200 500

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

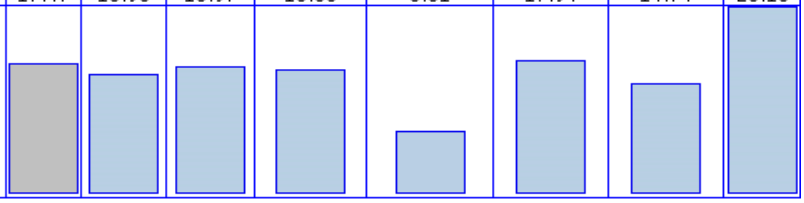
1	2012	ACAD	TechEngineerTeacher	A	B	C	, hands-on, engaging contexts in which to learn. The integration of concept mapping **helps teachers to focus** on the learning outcomes assoc
2	2012	ACAD	TechEngineerTeacher	A	B	C	activity with lower stakes attached that enables students to identify what they have learned and **helps teachers to identify** any possible gaps
3	2012	ACAD	PracticeNurse	A	B	C	long-term obesity. Research suggests that appropriate diet and exercise during the post-natal period may **help women to lose** their postnatal
4	2012	ACAD	PracticeNurse	A	B	C	Co-creating Health'. * What is the MAGIC project? * How might you **help patients to consider** the risks and benefits of a particular course of
5	2012	ACAD	PracticeNurse	A	B	C	their daily workload. This article aims to give an overview of the test, **helping nurses to understand** the procedure and its use in the primary
6	2012	ACAD	OccupationalHealth	A	B	C	www.oliverbrennan.co.uk Online resources for your research needs * www.mind.org.uk # Mind is a charity that **helps people to take** charge
7	2012	ACAD	TeachingExceptional	A	B	C	students to fulfill class assignments asynchronously. Utilizing assistive technology such as screen readers can **help students to process** large

Figure 18: The Concordance for <help><N><to><Vinf>

Figure 18 suggests that the structure is present in actual language use. Second, the structure is used in the academic setting, but is lower in frequency (672). 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. However, the frequency of the first pattern (as compared to the second one) is high. Hence, this <help><N><Vinf> sequence seems to be more preferable in academic writing. Further, the following Figure 19 shows the comparison of the two structures in BNC.

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	1682	159	270	121	87	296	226	523
PER MIL	17.47	15.96	16.97	16.66	8.31	17.94	14.74	25.10

SEE ALL SUB-SECTIONS AT ONCE



SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	3210	306	555	316	415	489	250	879
PER MIL	33.35	30.71	34.89	43.51	39.65	29.65	16.31	42.19

SEE ALL SUB-SECTIONS AT ONCE

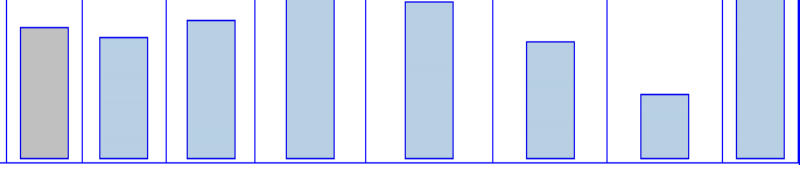


Figure 19: Help you to do VS Help you do in BNC

In BNC we can see that the two structures are in use both in academic and spoken sections. The occurrence in spoken and academic section is almost equal for the first structure, while for the second structure, the occurrence is relatively lower in academic section than in spoken section. The second structure differs in a way that there is a striking difference between non-academic (489) and academic writing (250).

MWU 2: Analysis

Similarities are observed from corpus validation. First, corpora investigation to both BNC or COCA suggests that both structures (with or without 'to' infinitive' are present in actual use (see Figure 17-19). However, COCA and BNC differ in terms of domination in academic domain. A closer observation can help us understand how COCA data suggests that the use of the structure without to infinitive is more frequent (see Figure 17 and 18). BNC data, however, is interesting as the frequency of the structure with 'to' infinitive is almost equal in academic (226) and non-academic (296) domains. Significant difference, however, is observed for structure without 'to' (489-250). That can be seen in Figure 16.

Therefore, we can safely assume that domain shift is relevant case-per-case. Studies in terms of the difference of spoken/written English, such as Biber (2006), and Carter & McCarthy (2006), or academic/non-academic English, like Allen (2009) and Annelie & Erman (2012), cannot be generalized as there are a number of lexis or structures applied in more than one domains. Even when we notice Coxhead's (2000) academic wordlist, which was later refined by Davies (2014) in COCA, some words are treated equally in spoken/written domains, such as negative (S2/W2). COCA data however, provides an interesting historical finding that the use of the structure with 'to' steadily increase over years (see Figure 17), while the use of the structure without 'to' is constant despite remaining higher in frequency.

MWU 3: Corpus Data

L1 seems to interfere when the vocabulary is similar to L2, not always on the lexis but it can also apply on the grammar. This happens to <goal>, in which its Indonesian equivalence is *gol*. Negative transfer from students' L1 affect certain lemmas with similar surface forms as their Indonesian translation. Clear-cut distinction of the lemma <goal> is observed between the two tasks. In the abstracts collection, the lemma is used completely in the sense of goal as an aim or objective. This is because the semantic of <goal> in Indonesian (*tujuan*) is not of the same equivalence. This might be the reason why students preferred to use *purpose* and *intention* instead of *goal* to express the same concept.

Table 6. Expressions for Aim and Score

Sense	Lemma (frequency)	
	Scientific Writing	Free Writing
Aim	*purpose (51), aim (19)*	*purpose (11), intention (7)*
score	-	*goal (39)*

Another sense of <goal> may also refer to points scored by team players in a sport game. In this case, the equivalence of <goal> in Indonesian, *goal*, is similar in form. This is the sense that is also present and dominating in the free writing. It is interesting to observe some verbs that collocate with <goal> as it suggests the interference of L1. Number --1,2,3,4, refers to the left context <make, create, score, print> respectfully:

1	for sure. But making	**goal**	in Chelsea was not easy for Torres, as ----------	1
2	very amazing! He created	**goal**	only in his first 5 minutes after the recovery ----------	2
3	while they have the same	**goal**	and equally work hard	
4	made. She had her private	**goal**	Without this effort, everything is	
5	are very passionate to	**goal**	to equalize the match, avoiding Atletico from	
6	Sturridge who has scored 19	**goals**	and 4 assists ----------	3
7	again? What made the	**goal**	of this movie seems----------	1
8	for sure. But making	**goal**	in Chelsea was not easy for Torres, as ----------	1
9	everytime a player print	**goal**	it is rewarded in Champion League ----------	4
10	"Dortmund will not make	**goal**	with time" Mourinho said as he was ----------	1

Figure 20: Concordance 5: The Retrieval of <goal> in Free Writing

I listed four verbs that collocate with <goal>. The first one, and the highest in frequency is <make> (line 1, 7, 8 and 10), <create> (line 2), <score> line 6, and <print> (line 9). One pattern in line 7 is not valid as it is grammatically wrong (*goal* is used as a verb).

MWU 3: Corpus Validation

The pattern <V><goal> is interesting to research as students have the tendency to choose verbs under the influence 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 collocate with <*gol*>, which are: <*mencetak*>, <*membuat*>, <*bikin*> and <*menciptakan*>. In this respect, <*membuat*> and <*bikin*> are similar; these verbs can be translated literally as the followings: 'to print', 'to make', 'to make (informal), and 'to create', respectively. See Figure 21.

The first part of Figure 21 from SEAlang Indonesian corpus (Scannel, 2010) has shown that there are three verbs that collocate with <gol> in Indonesian. The literal translation of the verb <score> in pattern 3 is not observed in this corpus. The three verbs that have been mentioned previously were often preferred by the students, but the strings resulted by the collocation patterns are odd; they are merely the concatenation of the literal translation of those verbs. In this case, the verb <score> is the perfect collocate to <score>. This hypothesis can be validated by retrieval on COCA as the standard corpus of English. See Figure 22.

SEAlang Corpus	*mencetak,membuat, bikin*	
setelah Gerd Muller mencetak	**gol**	dua menit menjelang turun minum
setelah David Trezeguet mencetak	**gol**	emas penentu kemenangan pada
Rummenigge dan Rudi Voller mencetak	**gol**	pada menit ke-74 dan 80.
ai tersebut pula, Maradona membuat	**gol**	yang sangat buruk pula.
setelah Lopez tak kunjung membuat	**gol**	Bermula dari bola liar yang
Maradona membuat	**gol**	terbaik sepanjang masa yaitu ketik
ersebaya, M. Afif, dan menciptakan	**gol**	tunggal untuk timnya. Ketidaklogi
lagi Engkau main tangan, bikin	**gol**	ke gawang salah satu dari kami yang
ia bisa bikin	**gol**	tanpa ada penjaga gawang
yang suka bikin	**gol**	sembarang tendang. "Ada uang, Tuan
Pan Localization Corpus	*membuahkan, membuka, mementahkan, menambah, mencetak, membukukan*	
-14 yang nyaris membuahkan	**gol**	namun berhasil digagalkan oleh Cerny
. Pemain depan Argentina itu membuka	**gol**	awal dari titik penalti pada
turnamen Piala Nasional Afrika, membuka	**gol**	awal pada babak pertama dan
Milan sedangkan United membukukan	**gol**	menit akhir lewat Carlos Tevez
di Nou Camp ketika mereka mementahkan	**gol**	pembuka Xavi dari titik penalti
Frederic Kanoute menambah	**gol**	bagi Sevilla pada malam yang
pertama tetapi Sevilla kembali menambah	**gol**	mereka mendekati turun minum,
ke-76 dan menambah	**gol**	keduanya dua menit menjelang pertandingan
dalam pertemuan kedua itu, mencetak	**gol**	kemenangan pada menit ke-
asal Venezuela Arango mencetak	**gol**	dari jarak dekat untuk menempatkan
Valladolid Sisi untuk mencetak	**gol**	kemenangan pada menit terakhir ditepis
usai ketika Hosni mencetak	**gol**	keduanya melalui tendangan sejauh 2

Figure 21: Comparison of verbs specifying <gol> in Indonesian Corpus

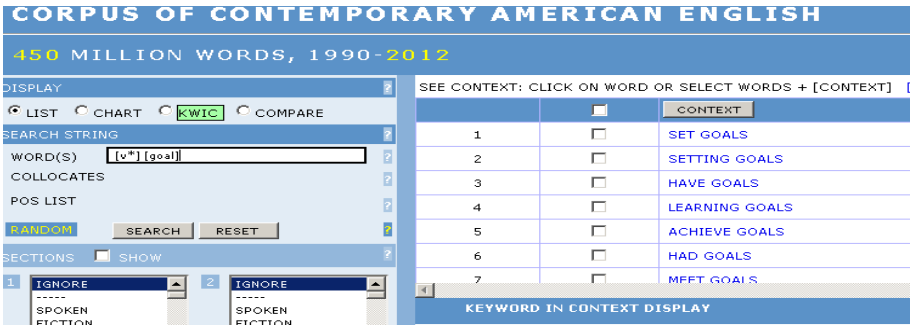


Figure 22: Concordance < V><goal>

Figure 20 describes verbs that do not necessarily relate to <goal> in the senses of points earned. The top six verbs (set and setting are two tokens of the same verb) do not seem to be the verbs that suffice the sport-definition of goal. However, this corresponds to the frequency. Figure 23 explains that <goal>, as 'aim/purpose', is frequently used more in the genre of 'academic' as compared to others.

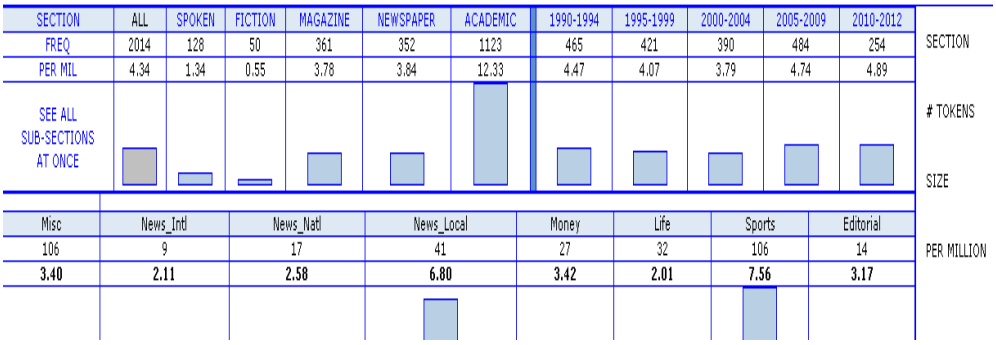


Figure 23: The Distribution Chart of <V*><goal>

In order to understand what verbs collocate with <goal> in the sense of scores, I refined the search to magazine section (Figure 24) specifically to the sport page (Figure 25).

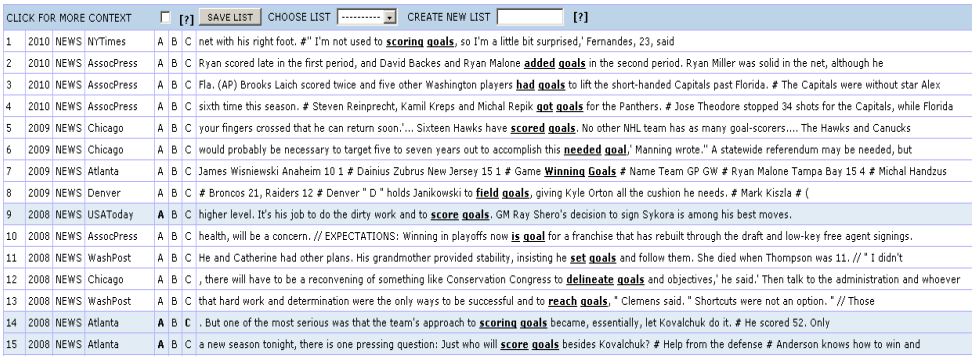


Figure 24: <V*><goal> in Magazine Corpus

News/Opin	Financial	Sci/Tech	Soc/Arts	Religion	Sports	Entertain	Home/Health	Afric-Amer	Children	Women/Men
60	31	37	21	16	51	9	61	35	27	51
3.23	**5.90**	**2.93**	**2.74**	**3.75**	**4.71**	**2.21**	**3.82**	**9.63**	**16.52**	**5.23**

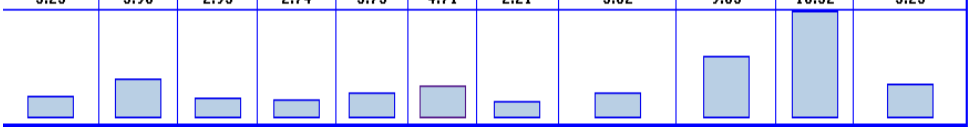


Figure 25: <V><goal> in Sport Section of Magazine Corpus

This refined search seems to generate positive result. As we specify closely, we begin to understand that the pattern <score><goal> is the most frequent pattern, which refers to the sense that we are looking for under sport domain within magazine section.

MWU 3: Analysis

Corpus data and corpus validation verify my proposition that the preference of other verbs such as <creates, make> is under the interference of students' L1, in this case Indonesian. It is also parallel to the findings of Husein and Mohammad (2012), Annelie and Erman (2012), and Arabski (2006) that concerned the negative impact of L1 to L2.

Although this negative transfer is common, to some extent it is dangerous when the choice of structures and vocabulary are considered peculiar. For peculiar use of the lemmas <create, make>, the readers perhaps can still deduce the meaning contextually. Domain shift is a common phenomena of metaphor study (Deignan, 2006). However, using <print> as a collocate to <goal> in terms of sport does not make any sense in English as <print> is a creation of textual image. It is true that some domain shift in metaphors might be shared across languages (Deignan & Potter, 2004), but *cetak* <print> and *gol* <goal> is a common metaphor in Indonesian but not in English. The shift might confuse native speakers of English.

CONCLUSIONS AND SUGGESTIONS

This research concludes that 1) students involved in this research were more exposed to spoken English rather than written English; 2) degree of exposure, as well as the students' L1, has a great influence on their preference over

certain structures; and 3) although the degree of exposure varies, teacher's instruction is still prioritized.

Considering the importance of teachers' instruction in this research, I recommend that teachers, especially in colleague-level academic writing class, instruct the students to consult different sources proportionally, and use the reference appropriately in accordance to the aim of their writing. Students cannot just rely completely on the teachers, or undergraduate final projects. I also strongly suggest that various corpora be used in the classroom in order to show authentic evidence of how language is used academically. Teachers can navigate the corpora together with the students in the class. The students are also encouraged to share the result of corpus exploration with teachers and other students. At this stage, teachers can always facilitate them in improving corpus exploration techniques, correcting false conclusions, or suggesting that they use a more specific/general corpus to explore.

REFERENCES

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS Learner Corpus. *Komaba Journal of English Education*, *1*, 105-127.

Annelie, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes 31*(2), 81-92.

Anthony, L. (2006). Concordancing with AntConc: An introduction to tools and techniques in corpus linguistics. *JACET Newsletter*, 155-185.

Arabski, J. (2006). Language transfer in language learning and language contact. In J. Arabski (Ed.), *Cross-linguistic influences in the second language lexicon* (pp. 12-21). Clevedon: Multilingual Matters.

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA.* Birmingham: Capstone.

Bautista, M.-S., & Gonzalez, A.-B. (2008). *The handbook of world Englishes.* New York: Blackwell.

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5(2)* , 97-116.

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers (Vol. 23).* Amsterdam: John Benjamins Publishing.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide; Spoken and written English grammar and usage.* Sprachen: Ernst Klett.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30-49.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly*, *34*(2), 213-238.

Davies, M. (2008). *American corpus*. From The corpus of contemporary American English (COCA): 385 million words, 1990-present. Retrieved from http://www.americancorpus.org.

Davies, M. (2014). *British National Corpus*. Retrieved from Corpora Collection: http://corpus2.byu.edu/bnc/?r=y

Deignan, A. (2006). *Metaphor and corpus linguistics.* Amsterdam/Philadephia: John Benjamins.

Deignan, A., & Potter, L. (2004). A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics*, *36*(7), 1231-1252.

Fauziati, E. (2003). Interlanguage errors in english textbooks for junior high school students in Surakarta. *TEFLIN Journal*, *14*(2), 21-33.

Figueredo, L. (2006). Using the known to chart the unknown: A review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing, 19*(8), 873-905.

Halliday, M. (1989). *Spoken and written language.* Oxford: Oxford University Press.

Harwood, N. (2005). Inclusive and exclusive pronouns in academic writing. *Applied Linguistics*, *26*(3), 343-375.

Hasan, R. (1986). *Grammatical cohesion in spoken and written English.* London: University College, London (Communication Research Centre).

Husein, A.-A., & Mohammad, M.-F. (2012). Negative L1 impact on L2 writing. *International Journal of Humanities and Social Science*, *1*, 184-195.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21.

Isaac, M.-F. (1986). *French creole interference in the written English of St. Lucian secondary school students.* (Unpublished MPhil Thesis, University of West Indies, Cave Hill).

Krauthamer, H.-S. (1999). *Spoken language interference patterns in written English.* New York: Berkeley Insight.

Kwary, D.-A. (2013). *Indonesian high frequency words*. Retrieved from The Indonesian High Frequency Word List (Version April 2013), http://www.kwary.net/iwl.html

Moehkardi, R.-D. (2012). Grammatical and lexical English collocations: Some possible problems to Indonesian learners of English. *Humaniora*, *14*(1), 53-62.

Nuruzi, M., Farahani, A., & Farahani, D. (2012). Deverbal Nominalisations across written-spoken dichotomy in the language of science. *Theory and Practice in Language Science 2*(11), 2251-2261.

Pang, W. (2010). Lexical Bundles and the Construction of an Academic Voice:A Pedagogical Perspective. *Asian EFL Journal*, *47*, 1-13.

Paumier, S. (2008). *Unitex manual.* Paris: Universite Paris Est Marne La Valee & LADL.

Roni, R. (2006). The students' competency in writing descriptive paragraph at Electrical And Mechanical Department, Faculty of Engineering, Tridinanti University Palembang. *TEFLIN Journal*, *17*(1), 28-35.

Sawalmeh, M.-H. (2013). Error analysis of written English essays: The case of students of the preparatory year program in Saudi Arabia. *English for Specific Purposes World*, *14*, 32-57.

Scannel, K. (2010). *Sealang*. From Indonesian text corpus: http://www.sealang.net/indonesia/ corpus.htm

Šimčikaitė, A. (2012). Spoken discourse markers in learner academic writing. *Studies About Languages*, *20*, 24-48.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Szmrecsanyi, B., & Hinrichs, L. (2008). Probabilistic determinants of genitive variation in spoken and written English. In T. Nevalainen, *The Dynamics of Linguistics Variation: Corpus evidence on English past and present.* (pp. 291-309). Amsterdam: John Benjamins.

Wijaya, D. (2012). Teaching English generic nouns: The exploration of the generic idea in English and Indonesian and the applications of explicit instruction in classroom. *Indonesian Journal of English Language Teaching*, *8*(1), 98-115.

Yembise, Y. (2011). Linguistic and cultural variations as barriers to the TEFL settings in Papua. *TEFLIN Journal*, *22*(2), 201-224.

Yoon, H.-S. (2008). More than a linguistic reference: The infuence of corpus technology on L2 writing. *Language Learning and Technology*, *12*(2), 31-48.