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 dictionaries. 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>](image)

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](image)

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](image)
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.

![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.

<table>
<thead>
<tr>
<th>SECTION</th>
<th>All</th>
<th>Movies</th>
<th>Lecture</th>
<th>Social Media</th>
<th>Assignment</th>
<th>Books</th>
<th>Magazines or Newspapers</th>
<th>Journals</th>
<th>Theses, Dissertations</th>
<th>Academic Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1900</td>
<td>421</td>
<td>15.9</td>
<td>42.0</td>
<td>13.46</td>
<td>11.79</td>
<td>75.69</td>
<td>75.73</td>
<td>62.28</td>
<td>182.81</td>
<td>135.76</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Written</th>
<th>Spoken</th>
<th>Academic</th>
<th>Non-Academic</th>
</tr>
</thead>
<tbody>
<tr>
<td>31%</td>
<td>69%</td>
<td>37%</td>
<td>63%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
<th>Type</th>
<th>Frequency</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>481</td>
<td>in</td>
<td>188</td>
<td>English</td>
<td>110</td>
</tr>
<tr>
<td>And</td>
<td>219</td>
<td>study</td>
<td>177</td>
<td>Foreign</td>
<td>99</td>
</tr>
<tr>
<td>Be</td>
<td>207</td>
<td>language</td>
<td>151</td>
<td>students</td>
<td>97</td>
</tr>
<tr>
<td>but</td>
<td>111</td>
<td>need</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).

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.

![Figure 8: The 3-Grams for Left-Right Context <suggest;help;goal> in Abstracts Collection](image)

### 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>
<thead>
<tr>
<th>Positive (28%)</th>
<th>Negative (72%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6%</td>
<td>Strong belief in the source</td>
</tr>
<tr>
<td>10%</td>
<td>Strong belief in frequency</td>
</tr>
<tr>
<td>12%</td>
<td>Strong belief in teacher’s instruction</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>MWU 1</th>
<th>Suggest that you do</th>
<th>- Suggest you to do</th>
</tr>
</thead>
<tbody>
<tr>
<td>MWU 2</td>
<td>Help you to do</td>
<td>- Help you do</td>
</tr>
<tr>
<td>MWU 3</td>
<td>Score goal</td>
<td>- make goal</td>
</tr>
</tbody>
</table>

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.
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:

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

**Figure 11: Concordance of \texttt{suggest}<\texttt{PRO}<\texttt{to}><\texttt{V}>

Figure 11 shows that the first pattern is in actual use, but very low in fre-
quency. 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.):

**Figure 12: Concordance of \texttt{suggest}<\texttt{Pers.PRO}<\texttt{to}><\texttt{V}>

[Table and diagram images are not included in the text.]
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:

**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.
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

<table>
<thead>
<tr>
<th>Vocabulary Profile (Token of academic word list)</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scientific Writing Tasks</td>
</tr>
<tr>
<td>More than 60%</td>
<td>4</td>
</tr>
<tr>
<td>Less than 60%</td>
<td>61</td>
</tr>
</tbody>
</table>
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.

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

![Figure 15: Concordance 3: The Retrieval of *help* in Abstracts](image-url)
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 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:

![Figure 17: The Concordance for <help><N><to><Vinf>](image)

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.
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**

<table>
<thead>
<tr>
<th>Sense</th>
<th>Scientific Writing</th>
<th>Free Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim</td>
<td><em>purpose</em> (51), <em>aim</em> (19)</td>
<td><em>purpose</em> (11), <em>intention</em> (7)</td>
</tr>
<tr>
<td>score</td>
<td>-</td>
<td><em>goal</em> (39)</td>
</tr>
</tbody>
</table>

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:
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><goal> from Indonesian corpus indicated the same result. There are four verbs in Indonesian that collocate with <goal>, 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 <goal> 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.
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 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


Prihantoro, Influence of L1 and Spoken English on English Writing  243


