Corpus-based Analysis of Collocational Errors

Myung-Jeong Ha

Department of English Language and Literature, Sangmyung University, Korea
mjha@smu.ac.kr

Abstract

The present study investigates the use of high-frequency verb collocations in L2 learner English. Although it has been widely accepted that collocations are essential to language use, until recently, the complexity of collocations was mainly hidden. However, developments in the new technology of corpus linguistics and the availability of large electronic texts from spoken and written sources have made possible new perspectives to learner language use. In addition, learner corpus data constitutes a new type of performance data which offers a valuable view of learners’ errors and non-errors as well. This study is thus embedded in a combination of elaborate manual investigation and computer-aided analysis using a concordance program, WordSmith Tools. The results of this study show that errors are not entirely mismatch between the verb and the noun, and other types such as prepositional errors and determiner errors were relatively more frequent among advanced L2 learners. It is also found that the interference of the Korean learners’ L1 plays a role in the use of collocations. Based on the results, some suggestions are made for collocation learning that is helpful for EFL learners.

Keywords: Computer learner corpora, Concordance program, Error analysis, Collocation

1. Introduction

In recent years, there has been an increasing interest in collocations as an important aspect of knowledge for second language learners [1]. Collocations are an important part of language use and collocational proficiency often differentiates native and non-native speakers [2]. Knowledge of collocations makes second language learners proficient and fluent, and thus makes a clear distinction between advanced language learners and intermediate language learners. Laufer and Waldman [3] observed that "lack of collocational knowledge may impede the comprehensibility of learners' expression. (p. 648)"

Although it is generally accepted that collocations are indispensable and problematic for foreign language learners and they therefore should play an important role in second language acquisition (SLA), L2 learners’ difficulties with collocations have not been discussed in detail by EFL practitioners so far [4]. Collocations have been largely neglected by researchers, material designers and EFL practitioners.

The computer-assisted analysis of corpus data opens up a new perspective for collocation analysis since linear text analysis is not necessary any more. Modern corpus-linguistic software packages like WordSmith Tools offer wordlists, concordancers, and other tools for analyzing the co-text of the investigated words for recurring patterns of use. The introduction of computer learner corpora and corpus analysis techniques into second language research allowed researchers to conduct studies that involved a large amount of language samples. In the field of computer-assisted language learning with new corpus technology becoming available, the need for more studies on collocations is obvious.

Following the trend towards collocational competence in second language learning, the present study investigates the use of English verb collocations in the writing of native speakers of Korean. Since restricted collocations are frequently said to be an area where L2 learners have greater difficulties but remains neglected, the use of restricted collocations in a learner corpus are mainly investigated. In spite of some suggestions made on the teaching of collocations in recent years [5], it is not clear how and which of collocations in a second language should be taught. To provide some answers to these questions, it is important to examine the difficulties that L2 learners have in using collocations.

1 This research was supported by a 2013 Research Grant from Sangmyung University.
Since the production of collocations is more problematic than the comprehension, the present study focuses on collocational problems in L2 learners’ writing in order to identify the difficulties they have. The purpose of this study is to explore aspects of how high-frequency common verbs in a learner corpus form collocations with other words. The analysis aims to illustrate the way corpus-informed descriptions can be used to increase the understanding of collocational difficulties of Korean EFL learners.

2. Defining Collocations and Computer Learner Corpora

Under this study, the term collocation is employed in a phraseological sense, meaning that it is used to denote a type of word combination rather than the co-occurrence of words in a certain span [6]. Following the most widely accepted defining criterion for collocations, or “arbitrary restriction on substitutability” (p. 225) [4], I take collocations to be associations of more than two lexemes where the combination is semantically transparent, but includes an arbitrary choice of more than one lexeme. Nesselhauf’s definition is supposed to make the task of distinguishing collocations from free combinations on the one hand and idioms on the other hand easier and more reliable.

Collocations are placed on the continuum between free combinations and idioms. A distinction is made between combinations in which a possible restriction on the substitutability of elements. Three types of word combinations can be classified in terms of the notion of restricted sense (e.g. [4][6][7]): (a) free combinations which include the verb and the noun are used can be freely combined (e.g. find an entrance); (b) collocations among which the sense of the verb is restricted and can be combined with certain nouns (make a suggestion but *make a determination); (c) idioms in which the verb and the noun are used with a restriction and the substitution is not acceptable/allowed (e.g. make up for).

In addition, collocations can be divided into two categories according to the word class of their constituents [7]: lexical collocations and grammatical collocations. While lexical collocations combine two open class words such as verb and noun, or adjective and noun, grammatical collocations combine an open class word and one closed class word such as verb and prepositions or a grammatical structure such as infinitive or clause. Under the present study, both lexical collocations and grammatical collocations are scrutinized.

Computer learner corpora have been in existence for a relatively short time. Computer learner corpora are electronic collections of spoken or written texts produced by foreign or second language learners [8]. Learner corpora are seen as a new resource for foreign language teaching researchers and educators [9-11]. Although learner corpus compilation is a new activity, a number of learner corpora already exist. Learner here refers to somebody learning a foreign language or to a foreigner learning the language in a country where it is spoken natively. Because it is characterized by a high rate of misuse such as lexical and grammatical errors, it is hoped that learner corpora will provide researchers with essential sources of controlled computerized data which can be analyzed with computer-aided error analysis. In particular, studies informed by learner corpora have revealed some of the main problems that EFL/ESL learners have when writing academic essays: errors involving the collocational patterning of words and phraseological infelicities.

The learner corpus used in the study is a Korean EFL learner corpus of writing essays. Given that learner corpus analysis has proved a useful tool to reveal a number of distinctive features of L2 writing discourse, the use of learner corpora will provide relevant information on the difficulty of particular collocational features from the learners' perspective, which can help to identify features which teaching should emphasize and to evaluate their difficulty. In addition, many studies of collocation have been corpus-informed while exploiting the potential of a corpus to identify linguistic categories.

3. Historical overview of collocation studies

Since the 1990s, scholarly discussions on the importance of collocations came to the attention of the field of second language education [12]. There are four primary reasons for the discussions:

First, TESOL practitioners began to criticize that second and foreign language teaching methods (e.g. Grammar Translation Method) overlook the role of lexicon. There have been greater needs for developing efficient ways to present vocabulary to second language learners [13]. Second, Lewis started implementing a collocation-based syllabus as well as the Lexical Approach [14]. Also, the easy

For these reasons, a great interest in collocation learning has emerged in second language acquisition. In fact, L2 vocabulary knowledge includes collocations and other types of phraseology. Multi-words expressions such as phrasal verbs, formulae, and collocations are highly significant for a native English speaker’s language competence and they are much more than lexemes [17].

Studies that examine L2 learners’ use of collocations are not sufficient. In 1993, Bahns and Eldaw [18] investigated German EFL students’ production of collocations consisting of a verb and a noun. Their approach involves devising experimental procedures for eliciting collocations from homogeneous group of participants. They reported that the number of collocation errors was twice as high as the number of single word errors in the translation tasks. In a study of the learners’ whose L1 is close to English, Bahns and Eldaw [18] argued that the learners’ mother tongue may lead to inappropriate collocational usages.

Similarly, Nesselhauf [4] investigated the writing of German EFL learners and reported that the most frequent error was wrong choice of verbs. She found that L1 had an influence on all types of collocational mistakes and that the collocations not consistent with learners’ L1 and L2 were much more difficult for learners to acquire. In another major study, Nesselhauf [19] extracted around 2,000 verb-noun collocations from the German Corpus of Learner English, demonstrating that a quarter of the contained errors and another third of the collocations were judged as erroneous. She reported that more than 50% of errors are induced by L1 influence. It is interesting to note that Nesselhauf’s recommendations include placing more emphasis on common collocations even at advanced levels and teaching how collocations are used in context and the ways in which they are restricted. In addition, Liu [20] examined collocational errors in 127 Chinese college students’ examination papers and their 94 compositions. She found that verb-noun collocation errors occurred most frequently. On the basis of the early error analyses and the recent collocation studies, many collocation errors are induced by L1 influence.

On another dimension, some authors (e.g. [21][22]) cast doubt on the emphasis on L1 transfer as the main cause of the collocational problems. In Wang and Shaw [22], different L1 groups made a similar number of collocational errors in their compositions (e.g., wrong combinations of nouns with do and make). In their analysis of collocation use in advanced learner English, Wang and Shaw [22] argue that the collocation side of language learning is more closely connected with syntax learning, which is usually affected by L2-based features. According to them, L1 transfer is only one of the factors.

In particular, Wang and Shaw [22] note that one type of collocational difficulties is connected with frequent, high-utility dynamic verbs on the basis of quite a few studies (i.e., [23-26]). Several studies reported L2 learners’ difficulties with collocations focusing on highly frequent delexicalized verbs. For example, Lennon [24] investigated the erroneous uses of put, go, recognize, and take and found that advanced learners of English showed lack of knowledge of collocational probabilities as to these high frequency verbs. Liu and Shaw [25] conducted a contrastive corpus analysis to observe the uses of high frequency verb, make, in learner and native writing. The analysis suggests that the learners’ writing appear repetitious due to lack of lexical richness.

In sum, the overall picture that appears based on collocation studies is that the use of collocations remains problematic for L2 advanced learners and that they have a great problem with the use of common words and collocations. Given that one of the collocational problems for advanced learners is associated with high-frequency common verbs, this study attempts to analyze the collocation problems of common verbs for advanced Korean learners.

4. Methods

4.1. Learner corpora

The corpus used in the study consists of 15,933 words, drawn from a larger corpus of 1,085,879 words. The larger corpus was created with a view to being enlisted to Learner Corpora Around the World (http://www.uclouvain.be/en-cecl-lcworld.html) developed by Sylviane Granger of the
University of Louvain. The larger corpus consists of argumentative and descriptive essays that were produced in English diagnostic tests at Yonsei university.

Yonsei English Learner Corpus was built up for the purpose of academic research. The 3,286 participants were Korean native speakers in the first semester as first-year college students in Korea. The essays were written under examination conditions and the students did not have access to dictionaries. Students were given a word limit of 100 - 300 words. The essays from the Yonsei English Learner Corpus to be included in the smaller corpus were selected to include those receiving C1 and C2 grades based on the 9 levels (i.e. A1, A1+, A2, B1, B1+, B2, B2+, C1, and C2). Given that collocations are an important part in English language teaching, especially at an advanced level, the focus of the study is put on the difficulties of advanced English learners with collocations. As a result, there were altogether 233 essays in the smaller corpus.

4.2. Data retrieval and analysis procedure

Overall data analysis procedure is shown in Figure 1.

Figure 1. Data Analysis Process

First, a frequency wordlist was generated by WordSmith Tools [27]. As mentioned earlier, I decided to choose high-frequency common verbs and their collocations for analysis. According to Altenberg and Granger [23], the following fifteen verbs are placed on any corpus-based list of high-frequency verbs: have, do, know, think, get, go, say, see, come, make, take, look, give, find, and use.

Given that the literature of high-frequency verbs points to an overuse of these verbs by EFL learners and the error-proneness of these verbs in L2 writing [24][28], I chose the node words based on these high-frequency verbs.

Next I obtained frequency scores for each verb lemma. In this paper, the composite set of words is viewed as lemma [3]. Figure 2 below shows an example of lemmatized results. To compute the verb frequencies in the corpus, I used WordSmith Tools’ lemmatizing facility, which enabled me to group all the inflectional forms of the lemma HAVE. In this way, the verbs in the corpus were lemmatized to get the total occurrences of each word.
Table 1 lists the frequencies of occurrence of the verbs used in the study. Of the fifteen verbs, I selected the most frequent verbs from the word lists of the learner corpus - that is, verbs that occurred more than 50 times in the corpus. In order to study the behavior of words in texts, it is necessary to obtain enough observations for each verb. It was thus difficult to deal with verbs with relatively low frequencies. All the necessary word-forms that are associated with the four key verbs were taken out and counted separately.

Table 1. Frequency of the fifteen high frequency verbs in L2 writing

<table>
<thead>
<tr>
<th>Verb</th>
<th>Occurrences</th>
<th>Lemmas</th>
<th>No. of Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find</td>
<td>3</td>
<td>find[1] found[2]</td>
<td>1</td>
</tr>
</tbody>
</table>

After choosing those verbs, I first obtained the frequency lists of each kind of verb form. The next step was to scrutinize concordance lines to weed out irrelevant instances. I then picked up collocation errors from the KWIC (Key Word in Context) lists with the help of KWIC index in WordSmith Tools. I classified them into the three categories, that is, free collocations, restricted collocations, and idioms. On the basis of previous studies [3][4], I used the BBI dictionary of English word combinations (BBI) and the Oxford Collocations Dictionary for Students of English (OCDE) as the main references, supplemented by the Collins Cobuild English Dictionary (CCED). As collocation errors were identified in the concordances, they were copied from the text and pasted into a separate file for later analysis.
5. Findings and discussion

Altogether 358 collocations were extracted from the learner essays, of which 140 appeared as problematic. Table 2 indicates that the highest rate of errors occurs in collocations of the node word, THINK (68%). The lowest rate of errors, on the other hand, is found with collocations of HAVE (35%). Not only the verb and the noun of a combination were judged, but also the non-lexical elements, such as prepositions and determiners, belonging to a combination, were judged unacceptable. Because some combinations contained more than two mistakes (e.g. in make a friend for make friends with a determiner and a preposition mistake were counted), the total number of mistakes in the 140 combinations was 170.

Table 2. Overall distribution of the High-frequency verb collocations in L2 writing

<table>
<thead>
<tr>
<th>Occurrence of verb collocations</th>
<th>Errors</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAVE</td>
<td>237</td>
<td>83</td>
</tr>
<tr>
<td>THINK</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>GO</td>
<td>58</td>
<td>25</td>
</tr>
<tr>
<td>MAKE</td>
<td>39</td>
<td>19</td>
</tr>
</tbody>
</table>

First, the results show that the advanced learners had a tendency to use wrong verbs which were similar in meaning. *Make problems, for instance, is a literal translation of the Korean expression “munje-reul mandeul-da”. Korean learners usually learn the meaning of make as “mandeul-da” and cause as “yagi-hada”. For Korean EFL students, the word make is more strongly associated with the phrase “munje-reul mandeul-da” (cause problems) than the word cause itself. This kind of error is caused by L1 transfer of verb meanings. The same type of errors occurred in the phrase such as *make serious accidents for “cause serious accidents”, *make side effect for “cause side effect”, or *make a smile for have a smile. Other types of errors were also identified such as *make relationships for “build relationships”. In the fields of ELT dictionaries, it is necessary to include usage notes on frequent errors such as *make problems under the entry cause. As for bilingual dictionaries, the dictionaries should incorporate information from L1 specific errors into the usage notes.

Second, the learner data also indicates that the learners make frequent errors in the use of prepositions or particles after verbs. In the case of THINK, non-lexical elements belonging to a combination, for example, *he always think first other person, were judged unacceptable. The example demonstrates that the preposition of a prepositional verb, here think, is missing.

As for the node word, GO, some students dropped the prepositions in phrases such as go into ... (*go army) or go on ... (*go trip)’. *Go army is literally translated from the Korean phrase “gundae-gada.” In addition, *go trip is a literal translation of the Korean expression “yeohaeng-gada.” In Korean, no prepositions are expected for these verb expressions, therefore this might be another case of L1 transfer.

Also it is confusing for the Korean students that the node word, GO, could be used without the prepositions if the following elements are adverbials or gerunds (e.g. *go to home for go home, *go to abroad for go abroad, *go to shopping for go shopping). These patterns appear as complicated for the learners and they have to learn the behaviors of each verb. It may be suggested that one reason for the EFL students' problems in learning English prepositions is that they usually try to learn the meaning and use of prepositions individually without paying sufficient attention to their collocational properties. Therefore, knowing which prepositions or particle should follow the verbs is another crucial part for Korean EFL learners. Assuming that the L1 influence on mistakes correlates with a high L1 influence on collocation production, these findings support claims of other researchers who argue that the interference of the learners’ L1 plays an undeniable role in the use of collocations (e.g. [4][18][19]).

Finally, except for the node word, THINK, all of the three verbs were associated with the problematic use of determiners. It indicates that the determiner mistake is evenly distributed over the essays.

Table 3 below summarizes the overall distribution of major types of errors.

Table 3. Distribution of major types of errors

<table>
<thead>
<tr>
<th>Occurrences</th>
<th>Examples</th>
</tr>
</thead>
</table>
"“Corpus-based Analysis of Collocational Errors1
Myung-Jeong Ha"
Determiner
Article or pronoun
missing, present though unacceptable

44

Thesedays I *go to the church with ... I watched TV show and *have a dinner I have a cold and I *had running nose we *had deep sleep

Preposition (verb)
Preposition of a prepositional verb
missing, present though unacceptable

28

Finally we *go to home. I *go to the shopping and moving. I *went to trip with my ... He always *think me

Verb
Wrong choice of verb

19

Finally I *had exam, it's ... he always *make a smile with friends have a picture that my mother *make he *makes a humor

First, determiner errors showing 44 occurrences were the most predominant. The result may be explained by the fact that there are no articles in Korean. Moreover, the use of English articles is a subtle and complex phenomenon, and there is no clear L2 input or formal instructions that can help Korean learners acquire the semantics of English articles [29].

Secondly, prepositional errors as in *go to home were the second most one, showing 28 occurrences. *Go to home is a literal translation of the Korean expression “jip-e ga-da.” It seems possible that these results are due to the fact that second language learners tend to put to use as a hypothesis that there is a word-for-word translation-equivalence between L1 and L2. That is to say, major difficulties appear to be L1 related.

Compared to the use of determiners and prepositions of a prepositional verb, fewer mistakes were made with respect to the type of wrong choice of verb. This finding is interesting because the verb in a collocation has a restricted sense, which makes its correct use more difficult.

In summary, overall the Korean EFL learners made relatively fewer mistakes of the type of wrong choice of verb, whether in the free or the restricted collocations. It seems reasonable to suppose that the Korean learners were more conservative and cautious in the choice of a noun collocate for a particular verb, whereas they tended to be more subjective and even creative in the choice of a preposition collocate for a certain verb, hence generating more mistakes in this respect.

6. Conclusion

The main conclusion that can be drawn from this study is that, to a certain degree, advanced learners still have a problem with the use of high-frequency common verb collocations. Although the students learned high-frequency verbs very early, once they have been taught, they tend to be overlooked.

As discussed earlier, the students lacked knowledge with respect to the collocational possibilities of verbs: there is a mismatch between lexical items as in to *make a smile. Whereas previous studies have mainly focused on the combinations of two lexical items, the present study reveals that verb collocational errors are not entirely mismatch between the verb and the noun. Rather other types such as prepositional errors and determiner errors were relatively more frequent among the advanced learners. It is therefore reasonable to highlight that the non-lexical elements belonging to a combination should not be neglected.

To some degree, their collocational errors can be explained as being the result of transfer from L1. In view of the influence of L1 transfer on the learners’ use of these high-frequency verbs, the present study is in line with Lennon’s [24] suggestion that “learners would benefit from consciousness raising as to areas in which lexico-semantic divisions do not correspond in L1 and L2. (p. 35)” Also the Korean learners’ reliance on L1 transfer on word combinations may indicate that they construct messages from individual words rather than prefabricated patterns [30]. This is consistent with Sinclair’s [31] argument that L2 learners are subject to the open choice principle more than to the idiom principle.

Several limitations to this study need to be acknowledged. The sample size is relatively small because this study was focused on the verb collocational errors from advanced Korean learners. And this study did not employ error annotation which is particularly relevant for interlanguage studies because POS taggers have been trained on the basis of native speaker corpora. Notwithstanding these limitations, the present study makes a noteworthy contribution to a growing body of literature on...
corpora of learner language of which insight is now beginning to gradually influence teaching syllabi in English as a second language. Interest in computer learner corpora is growing fast and has generated a number of studies, which emphasize the potential of this new resource for second language learning. This research will serve as a base for future studies of computer learner corpora that should eventually produce more reliable statistics on error distribution. For the field of CALL, the recent rise in interest in collocational competence in computational linguistics is seen as a welcome development and a great challenge. The growing knowledge about word combinations, together with the findings from this research, will eventually allow developers of CALL programs to go beyond the limitations of current commercial CALL software by starting the analysis with collocations and their contexts.

The results also have some pedagogical implications with respect to the teaching of collocations. Although rote learning has been criticized, it seems relevant to teach collocations explicitly. It is necessary to teach which collocate (the arbitrary part, e.g., make) is to be combined with the base (the non-arbitrary part, e.g., decision. The first step towards this is awareness that collocations differ from language to language. In addition, preemptive Focus-on-Form [32] may be useful for teaching collocations because it allows learners to raise their awareness of the difficulty of collocational use. Concordance-based exercises extracted from L1 corpora may act as a useful resource for learners’ awareness raising of the collocational use of high-frequency verbs.

Whereas studies of collocations have showed convincing results for the explicit teaching of collocation in the classroom, there still remains the issue of which collocations should be given priority and how they should be taught. The findings in this study shed light on this issue. As previously discussed, not all errors occurring are a mismatch between the verb and the noun that concerns the collocational possibilities of the two lexical items in question. Other types of errors such as prepositional errors as in thinks first other person (thinks of) and determiner errors as in go to the church with my friend (go to church) are also frequent among the Korean EFL learners. These results suggest that teaching grammatical collocations as well as lexical collocations for the frequent node words would be of great benefit to Korean EFL learners.

It is impossible to teach all of the collocations in a second language, we need criteria to choose which collocations should be included in teaching materials. Collocations should be clearly acceptable and frequent in a neutral register or any special register that is useful to the learner. For example in an academic writing course, combinations such as conduct a study or make an analysis would seem useful. Further research guarantees comparing the development of collocations from advanced EFL learners with that of collocations from low proficiency EFL students and will suggest instructional methods that are useful for learners with different levels of proficiency. In addition, the use of POS-tagged learner corpora should be encouraged toward more refined linguistic analysis.

7. References


