A Preliminary Experimental Study on Inherent Association of Verbs to Specific Nouns

I Gede Semara Dharma Putra[1),] Gede Primahadi Wijaya Rajeg[2)]

[1]Universitas Udayana, Denpasar, Indonesia
semaradharmaputra12@gmail.com
[2]University of Oxford, Oxford, England
primahadi_wijaya@unud.ac.id

Abstract

The research aims to analyse verb-noun collocations with different method. Collocations, or word co-occurrences, are typically studied based on written corpus data. In this paper, the method used is preliminary experimental method to explore the inherent association between a verb and certain nouns. We hypothesized that specific verbs would vary with respect to the nouns the verbs are associated with. To test this hypothesis, we developed an experiment using the Gorilla Experiment Builder to simulate the preference selection of nouns with each verb. The research was done with a sample size of 17 participants and each of them performed 35 trials inside the Gorilla software. The responses were compiled into a set of tables as the basis for generating bar plots, showing the frequency of nouns selected with each verb. The research results show that the studied verbs have certain preferences towards specific nouns. Even though this research found initial support for the hypothesis, the findings are not conclusive due to the small sample size of participants, being far from the population size initially measured. We discuss a way to corroborate the findings using different methods.

Keywords: association, collocations, nouns, verbs

1. INTRODUCTION

Corpus has played a significant role in providing researchers with a large body of text data that can be used to study frequencies and language usage patterns. Some studies have delved deep into the significance of corpora in the teaching of collocation (Chan & Liou, 2005; Daskalovska, 2015; Gablasova et al., 2017). Collocations can be defined as words that recur as a combination of two or more words of the same or different word classes (Ackermann & Chen, 2013; Shin, 2006 in Yamagata et al., 2023). Words that combine with each other do not always need to be adjacent to each other to be a collocation. As long as the sense is still present, the collocation will almost always be identified (Boers et al., 2014; Wood, 2020 in Yamagata et al., 2023). Collocation studies have been linked to L2 learning as it helps find the probability of students' usage in vocabulary choices, such as combinations of verb-noun collocations.

Most of the time, native speakers have an edge in the accuracy of verb-noun collocation selection than L2 learners because of native speakers' pre-defined acquisition which helps them determine specific nouns for specific verbs naturally (Altenberg and Granger, 2001 in Yoon, 2016). Research on the usage of collocations is still being done due to problems still being encountered from the viewpoint of L2 learners and that is why a substantial analysis is important to understand the spread of the learning process (Nesselhauf, 2003).

The research in this paper also aims to analyse verb-noun collocations/associations. However, we use a different method, namely an experimental method, to determine the degree to which a given verb is associated with certain nouns. Our paper, nevertheless, does not delve much into L2 learning as most researchers do. Yet, this research still puts itself in an area of L2 learners, simply not in a way that would create an impact directly on the L2 learners. The important thing being analysed here is how much of an agreement the L2 learners make following the verb-noun collocations.

Studies have been conducted regarding collocations. Church and Hanks (1990) outlined the basics of the calculations for collocations in a corpus using Mutual Information (hereafter MI) scores. They proposed the idea that the combination of words follows a specific set of norms, called "word association norms", which is measured using the MI. The research uses the Associated Press (henceforth AP) Corpus to find the collocations and analyse the MI scores. Because Church & Hanks (1990) mainly dive into a corpus, they can find various results from the calculation of their probability.

Aside from corpus-based calculations, other researchers dive more into language learning. Yoon (2016) conducted a longitudinal study comparing the phraseological competence between Native Speakers (NS) and Non-Native Speakers (NNS). The research is done by having these two groups of speakers compose essays of two types: argumentative and narrative, with NS composing two essays, and NNS composing six essays. In these essays, the collocations are then extracted and compared to a corpus specifically constructed by the researcher. Again, this research also used MI score to measure phraseological competence between the two groups. The results demonstrated a significant competency difference between NS and NNS writing of argumentative essays while generating little differences in the narrative essays.

Nesselhauf (2003) also researched the collocations from essays composed by L2 learners. Unlike the previous research, Nesselhauf (2003) investigates the difficulties that these L2 learners face on the collocation construction, based on the German sub-corpus of the International Corpus of Learner English (ICLE). The mistakes that the L2 learners made were then compared. The results showed a good number

of correct answers, with a handful of mistakes distributed among the category of free combinations, collocations, and idioms.

Another research by Ucar and Yükselir (2015) investigated the impacts of corpus-based activities. Types of corpus-based activities vary depending on the purpose, yet the fact is that all of them will use the corpora. An example of this is the use of corpora to aid in legal translation as it is important to know the correct words in a legal document (Nebot, 2008), another example is the encouragement of the usage of academic writing in students as a standard for higher-level education (Dong & Lu, 2020), and the last example is a test of identifying collocations which is the activity performed by Ucar and Yükselir (2015) in their research. The participants in this research are divided into two separate groups of participants: control and experimental, whether there is an effect on those activities in collocation teaching. The activity performed here is the activity of collocation tests, separated into a pre-test and a post-test. Both groups were treated to different conditions of approaching the activities beforehand and then directed into pre-tests and post-tests containing collocation tests. The tests were pre-determined using concordance lines from the Corpus of Contemporary American English (COCA) (Davies, 2010; for the audio-visual tutorials, see Rajeg, 2020). The results generated little to no significant differences between the pre-test and post-test scores within each group and the pre-test scores from both groups but generated a significant difference in the post-test scores from both groups.

Considering all the literature that has been reviewed above, this study investigates the distribution of noun preference for a set of verbs and their synonyms when participants are presented with a given verb and the nouns in each trial. We studied five groups of verbs (and their synonyms), namely *take*, *make*, *find*, *get*, and *give*. We use these verbs as the starting point because of two reasons. First, in COCA, these verbs belong in the top 100 most frequent words (see Figure 2 below). Second, they are generic verbs that any EFL learners of A1 level will learn. The hypothesis is that each of these verbs differs in its collocations with (i.e., preference for) the noun even though the verb and its synonyms are presented to participants with the same set of nouns.

2. RESEARCH METHOD

This paper adopted a mix of quantitative and qualitative methods. Quantitative analysis is used to determine the frequency of the noun chosen for a given verb stimulus. Meanwhile, the qualitative part of this research is used to interpret the results obtained from the research.

2.1 Participants

Since this study is based on a preliminary term-paper student project of the first author, with limited time constraints and difficulty to recruit a large pool of participants, we managed to only recruit seventeen (17) participants. The participants were students of the Bachelor of English Literature program at Udayana University. Given the small sample size, it is inevitable that there will be imperfections in the results, which nevertheless are sought to offer a preliminary trend. However, we will end the paper with how future studies can take the results forward using different methods.

2.2 Design and Material

The experiment was designed using the Gorilla Experiment Builder. The experimental design incorporates a multiple-choice question type of experiment. This means that the participants will be presented with verbs and noun options that the participants think are associated with the verb. The verbs are chosen based on their popularity/frequency ranking in the Corpus of Contemporary American English (COCA). This is explained below with reference to the "BROWSE/RANDOM" function in COCA (see Figure 1).
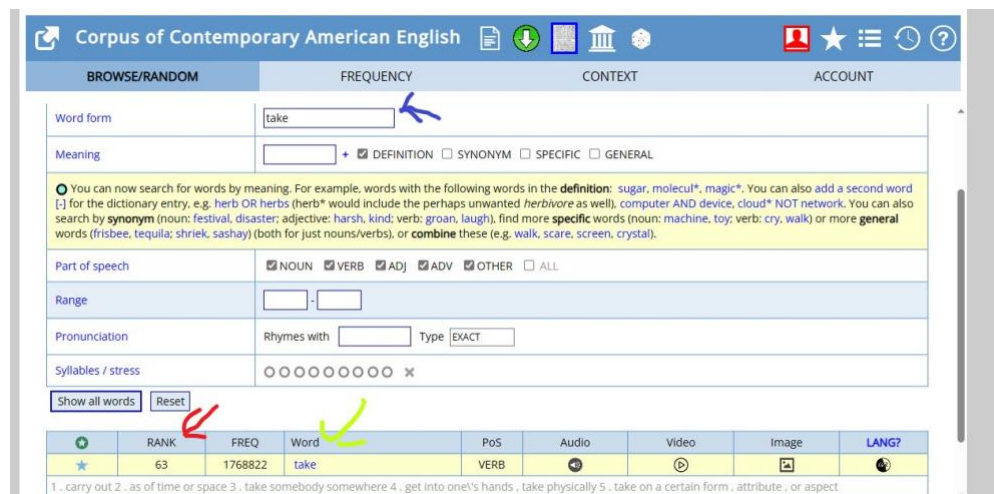

Figure 1 The "Browse" function in COCA.

Let us use the word *take* to determine its ranking in COCA. The blue arrow in Figure 1 points to the blank field where we can type the desired word. Once we click the "Show all words" button (to the left of the "Reset" button), a table below it will show up, showing frequency of the verb and its rank in COCA frequency list; the red arrow in Figure 1 indicates the rank for the verb *take* in the word frequency list in COCA. In this research, the studied verbs *take*, *make*, *find*, *get*, and *give* are among those within the top 100 most frequent words (see Figure 2).
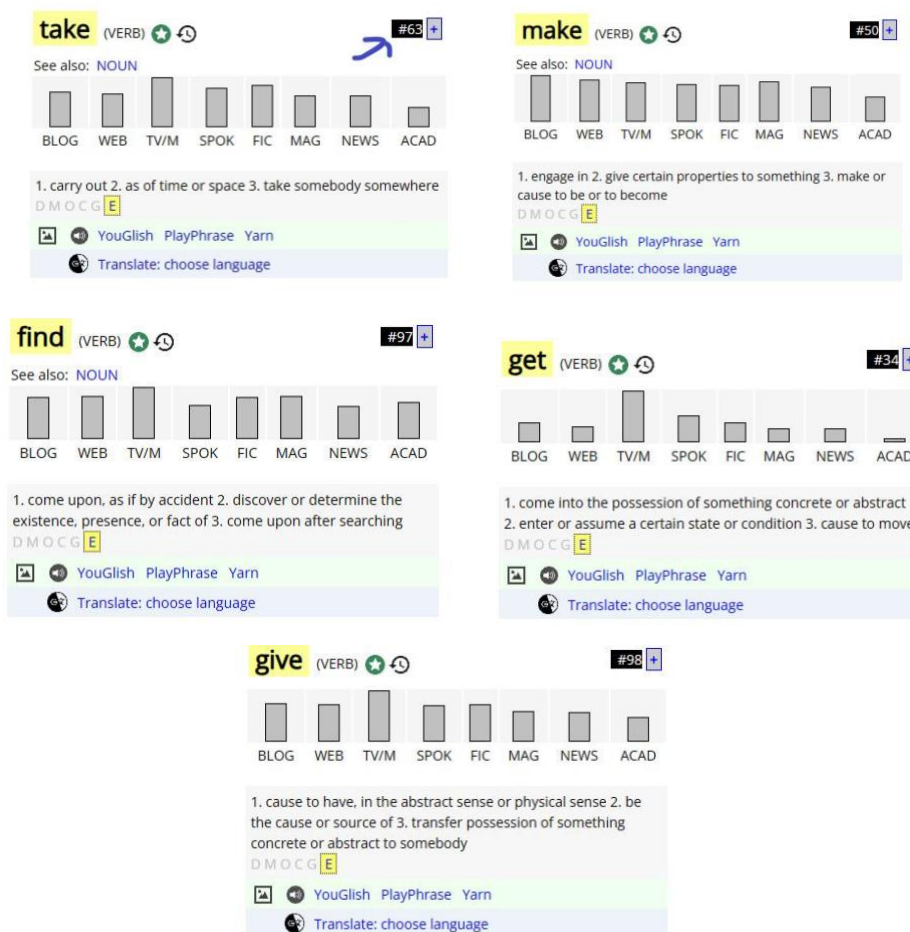
Figure 2 The ranking of the studied verbs in COCA (the ranking is pointed out by the blue arrow).

In addition to showing the ranking of the five central/main verbs in each group (ranking shown by the blue arrow in Figure 2), Figure 2 also shows the relative frequency of the verbs across different genres in COCA. The frequency ranking of the verbs was used to determine their order in the Gorilla experiment interface. In addition to these central verbs (i.e., *take*, *make*, *find*, *get*, and *give*), we also studied six synonyms of each of these verbs, making a total of seven verbs for each group. Their synonyms are all ranked outside the top 100, with the sixth synonym of each main verb being outside the top 10,000 words. Figure 3 illustrates the list of synonyms for *take* generated by COCA. The synonyms were chosen introspectively.

351

Figure 3 List of synonyms for *take* generated by COCA.

Unlike Ucar and Yükselir (2015) who used COCA to pick up pre-determined concordance lines, this research only used COCA to find the verbs and their popularity rankings. Meanwhile, the nouns were randomly picked concrete nouns. The nouns are picked consciously to reflect the illusion of one of the nouns being more common than others, and another being more uncommon than others. The nouns are not always strictly adjacent to the verbs and can be non-adjacent (Boers et al., 2014; Wood, 2020 as cited in Yamagata et al., 2023).

Figure 4 shows the design of the experiment in the Gorilla Experiment Builder, which has been given access publicly to readers (https://app.gorilla.sc/openmaterials/622948).
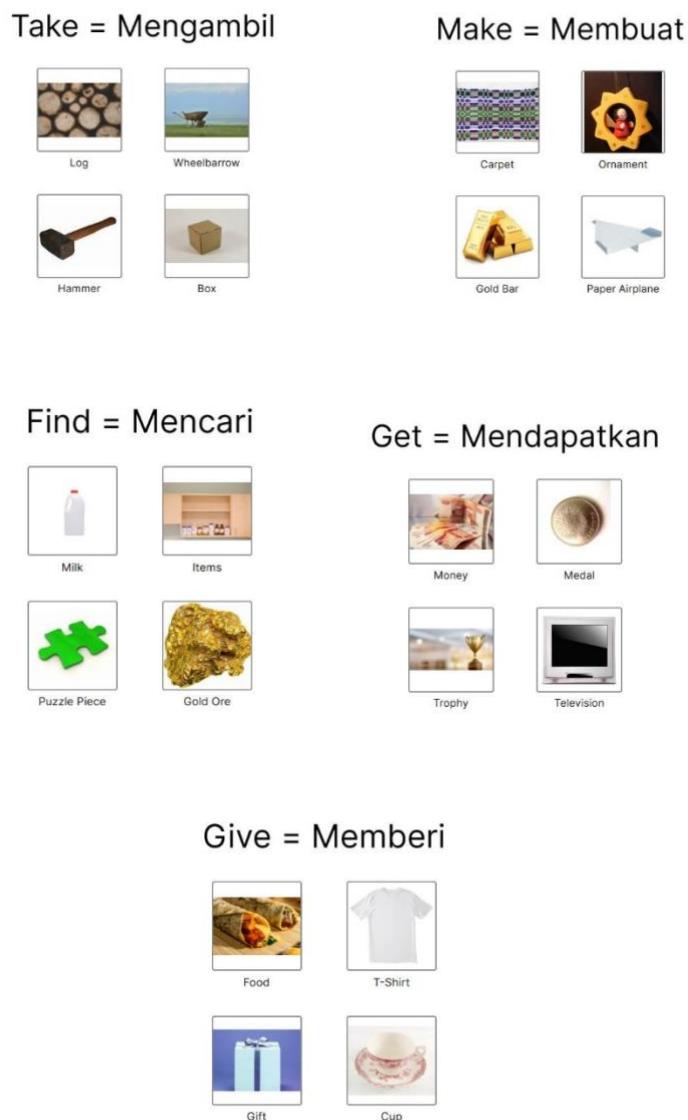
Figure 4 The design of the experimental trial.

Figure 4 shows the five main verbs with images of labelled object options (i.e., the nouns) under them (see Table 1 for reference). The goal of the trial is to determine which noun is selected for a given verb shown to the participant. The quantitative analysis quantified the frequency of the nouns chosen by the participant for a given verb. The English words also have an Indonesian translation to further stimulate their synonymy.

Table 1 The list of main verbs, their synonyms, and the nouns used for the experimental trial

Verbs	Nouns			
Take (main), Bring, Collect, Gather, Seize, Fetch, Hoard	Log	Wheelbarrow	Hammer	Box
Make (main), Create, Design, Produce, Establish, Construct, Manufacture	Carpet	Ornament	Gold Bar	Paper Airplane
Find (main), Search, Locate, Discover, Explore, Investigate, Rummage	Milk	Items	Puzzle Piece	Gold Ingot
Get (main), Receive, Earn, Acquire, Gain, Obtain, Procure	Money	Medal	Trophy	Television
Give (main), Share, Present, Provide, Grant, Donate, Bestow	Food	T-Shirt	Gift	Cup

2.3 Procedure

The experiment was conducted fully offline using the Gorilla Experiment Builder. Some of the 17 participants did the experiment on their chosen date and time. One day prior to the experiment, a training session was held to provide an overview of how the experiment will look like. On the next day, the participants met with the first author at a dedicated place to do the experiment, which lasted approximately less than five minutes per participant. Each participant received a total of 35 experimental trials (i.e., 5 groups of verbs * 7 verbs in each group). The participants were given up to 15 seconds to complete each trial to minimise the amount of time taken to do the experiment. Figure 5 shows the flow of the experimental trial with the verb *take* and its synonyms. During the trial, the image and the noun labels stay constant across the verbs to determine whether the noun selection is likely to be influenced by the different verbs.
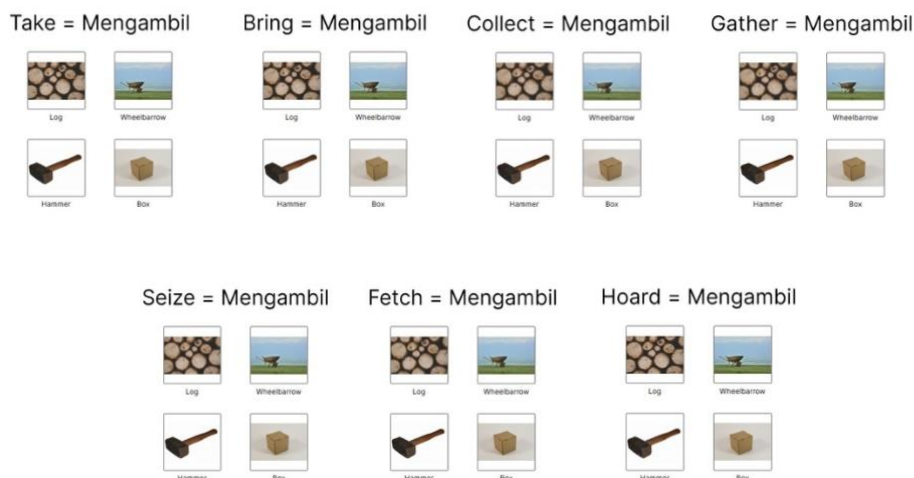
Figure 5 The flow of the trial for *take* and its synonyms.

From 17 participants, the experiment generated 591 tokens of data across each verb paired with specific nouns. It is important to note that there were four cases where the participant could not complete the whole 35 trials. The reason for this is that we set a limit of 15 seconds for each trial and the participants could exceed the time limit while thinking of an answer. The descriptive statistical analysis of frequency count and the visualisations were performed in R (R Core Team, 2020). The R codes for such analyses and the raw data are available at https://doi.org/10.17605/OSF.IO/SP4KB. The online R notebook presenting the R codes can be accessed at https://gederajeg.github.io/verb-noun-association/.

3. FINDINGS AND DISCUSSION

The experiment contains five main verbs (*take*, *make*, *find*, *get*, *give*) and six synonyms for each of those verbs (see Table 1 for reference). Once a participant clicked a picture for a given verb stimuli, the data is stored as a comma-separated-value (.csv) file. As there are five groups of verbs, the results will show five figures of statistical graphics (Figure 6-Figure 10) with seven bar plot groups of noun preferences for each verb. The distribution will be analysed to explore how favourable a noun is for a given verb. We will begin with the TAKE verb group visualised in Figure 6.
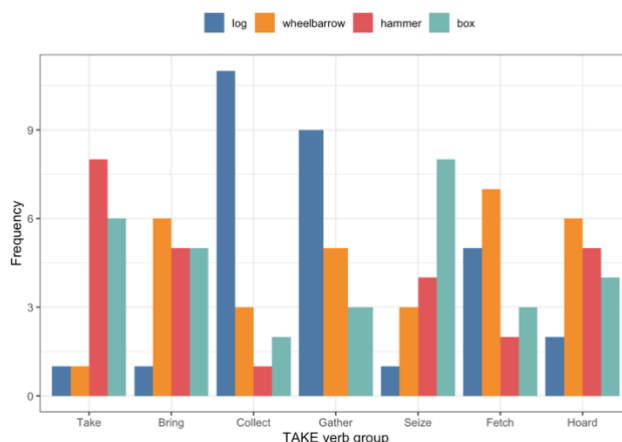
3.1 Results from the TAKE group

Figure 6 Frequency of nouns selected for verbs in the TAKE group

Figure 6 reveals that for each verb, nearly all nouns are chosen at least once. The exception to this is for *gather*, in which one of the four noun options (namely *hammer*) is not selected at all by the participants. Another trend that can be observed is that verbs can have different and similar noun preferences (i.e., operationalised as the frequencies with which the nouns are selected for a given verb); the examples are *fetch* and *hoard*. These verbs are similar in terms of their most preferred noun (i.e., *wheelbarrow*), but differ in the preference of the other nouns: *log* is more frequent for *fetch* than for *hoard*, while *hammer* is more frequent for *hoard* than for *fetch*. Moreover, *collect* and *gather* exhibit an interesting similarity in the sample. For both verbs, participants mostly chose *log* as the most preferred noun, though *wheelbarrow* is also relatively dominant for *gather* but not for *collect*.

As an example of contrast in the preferred noun, it can be seen between *gather* and *seize*. For the former, most participants chose *log* and *wheelbarrow* as the nouns they thought to be associated with *gather*. These nouns (*log* and *wheelbarrow*) in contrast are not judged by participants to be associated with *seize*, hence their low frequencies with *seize*. Instead, most participants chose *box* to be associated with *seize*. Further interesting contrast and similarity is shown by *take* and *bring*, which are semantically similar in expressing caused accompanied motion. The noun *wheelbarrow* is most preferred for *bring* but not for *take*, which in turn is predominantly associated with *hammer* and to a lesser degree with *box*. Yet, for *bring*, it seems to show no big distributional difference between *wheelbarrow*, *hammer*, and *box*.

Looking at the first set of results for the TAKE group, we can observe that most participants could converge in noun preference of a given verb.

3.2 Results from the MAKE group

The distribution of nouns for the MAKE verb group is shown in Figure 7. The nouns are *carpet*, *ornament*, *gold bar*, and *paper airplane*.
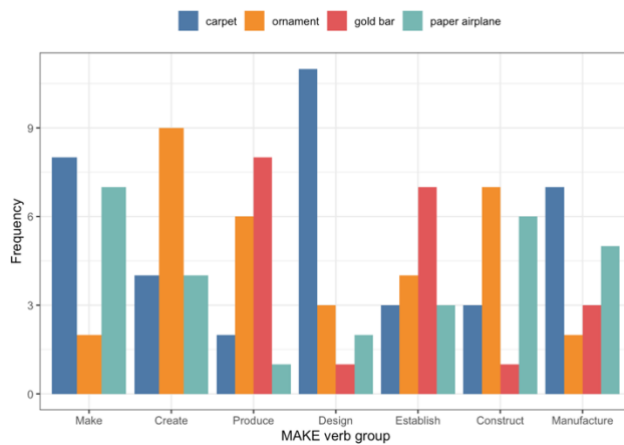

Figure 7 Frequency of nouns selected for verbs in the MAKE group

As for TAKE (Figure 6), we also see asymmetric preferences of the verbs with certain nouns. One clear example is shown by *make* and *create*. What is interesting about these verbs is that all participants never selected the noun *gold bar* to be associated with *make* and *create*. Perhaps, *gold bar* is too specific of a noun to be paired with semantically generic verbs such as *make* and *create*.

Furthermore, we can also observe that certain nouns are preferred for a set of verbs. For example, the verbs *make*, *design*, and *manufacture* can be conceived to be similar given *carpet* is their most frequent noun. However, unlike *make* and *manufacture*, the verb *design* almost has no preference for the *paper airplane*, which is, in turn, the second most frequent noun chosen by participants for *make* and *manufacture*.

Another interesting result is that *create* is the only verb that has a dominant preference for Christmas *ornament*. *Construct* is also similar to *create* in its main preference for Christmas *ornament*, but also for *paper airplane*, which is not the case in *create*. In a similar way, both *construct* and *create* have little to no preference to *gold bar*. An interesting thing to note is that *gold bar*, despite having a low preference in other verbs, is dominant on the verbs *produce* and *establish*. According to the results, most participants tend to think of the combination of "*produce gold bar*" or "*establish gold bar*" than other collocations.

3.3 Results from the FIND group

The distribution of nouns for the FIND verb group is shown in Figure 8. The nouns are *items*, *puzzle piece*, *gold ore*, and *milk*.
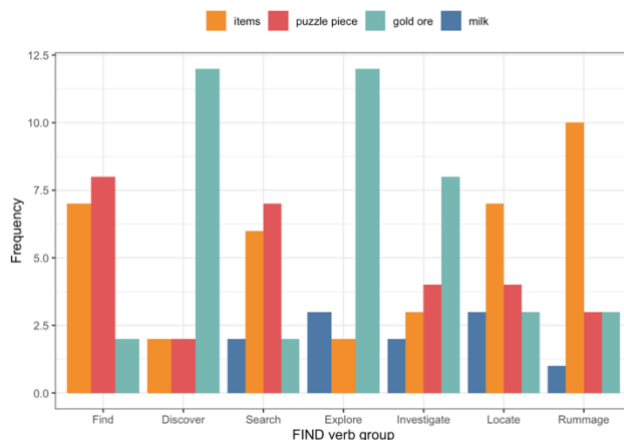

Figure 8 Frequency of nouns selected for verbs in the FIND group

Figure 8 primarily shows *gold ore* is predominantly selected for three verbs: *discover*, *explore* and *investigate*. The rest of the nouns in these three verbs were selected only about four times or less. The least-selected noun is *milk*, only 11 out of 118 times, and the noun *milk* is unselected at all in the verbs *find* and *discover* (see the non-existent blue bar in *find* and *discover* in Figure 8), although it enjoys some moderation in the word *explore*.

The verbs *find* and *search* almost have a similar preference to the nouns *items* and *puzzle piece*. Nevertheless, as the popularity ranking gets lower, *items* is selected more frequently than *puzzle piece*, especially in *rummage* where *items* is strongly preferred. Unlike the other verbs, the verb *locate* has a more balanced result while still leaning more in preference towards the noun *items*.

3.4 Results from the GET group

The distribution of nouns for the GET verb group is shown in Figure 9. The nouns are *money*, *medal*, *trophy*, and *television*.
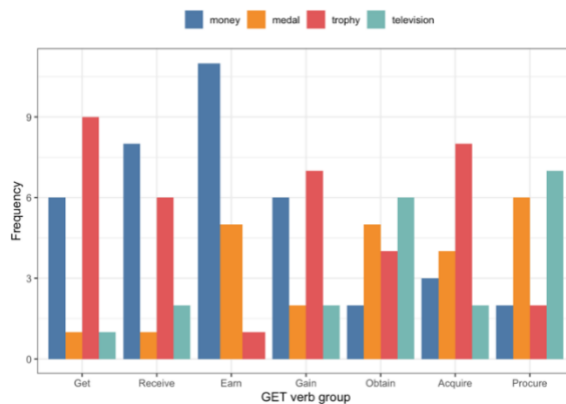
Figure 9 Frequency of nouns selected for verbs in the GET group

In Figure 9, we can see every noun was selected at least once, except for *television* which was unselected for the verb "Earn", suggesting that every participant agreed that *earn television* is unlikely. *Money* is dominant in the verbs *receive* and *gain*. However, unlike *earn* which favours *money* by a wide margin, *receive* and *gain* also have a similar preference for *trophy*, but these verbs do not have as much preference in *medal* as *earn*.

Trophy is more preferred in the verbs *get* and *acquire*. Although their preference number is the same, in the verb *get*, the number of instances of *money* is closely equal to *trophy*, unlike in *acquire* which is not that close. In *receive* and *gain*, *trophy* is closely tied together with *money*, which suggests that some participants may think "*receive money*" or "*receive trophy*", OR "*gain money*" or "*gain trophy*" according to the results.

An interesting part is that Television is more preferred in the verbs *obtain* and *procure*, although it is not as high in dominance as other dominant nouns among the other verbs. *Procure* also has an almost equal preference to *medal* alongside *television*, meanwhile in *obtain*, the difference is quite minimal, suggesting that *obtain* is quite widespread in its collocation usage according to the results.

3.5 Results from the GIVE group

The distribution of nouns for the GIVE verb group is shown in Figure 10. The nouns are *food*, *t-shirt*, *gift*, and *cup*.
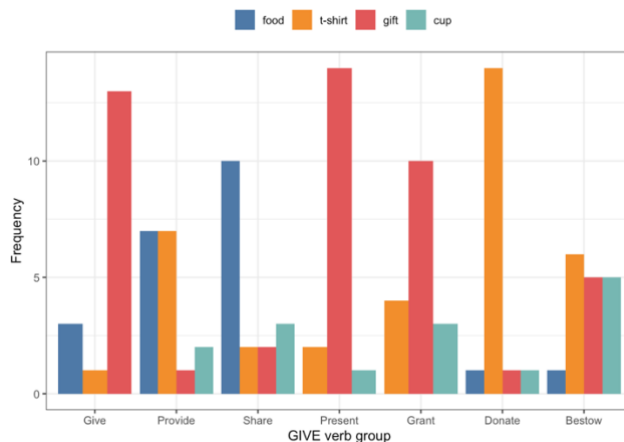
Figure 10 Frequency of nouns selected for verbs in the GIVE group

In Figure 10, we can first see that *cup* is not selected for the verb *give* and the noun *food* is unselected with the verb *present*; this suggests that the combination of *give cup* and *present food* are unlikely according to the result. The frequency of *food* gets even lower in the less-popular synonyms of *give*, except in *provide* and *share*. Although *food* is more frequent with the verb *share*, the noun is equal in preference to *t-shirt* for with verb *provide*, suggesting that both *provide food* and *provide t-shirt* are likely. Speaking of *t-shirt*, the only time it is predominant is with *donate*. This suggests that *t-shirt* is more likely to be donated than the other noun options.

On another note, the noun *gift* is much more preferred for the verbs *give*, *grant*, and *present*, suggesting that "*give/grant/present gift*" are typical collocation according to the results. Unlike the other three nouns, *cup* is the least selected among the three aforementioned nouns. It only has some moderate preference in *bestow*, and that verb does not have so much difference in its preference between *t-shirt*, *gift*, and *cup*, while in the previous six verbs, *cup* is selected only in small numbers of times.

3.6 Discussion

This preliminary experiment was designed to determine the noun preferences of verb synonyms based on the frequency with which each noun is chosen for a given verb by the participants. The results shown in Figure 6 to Figure 10 provide initial support to the hypothesis that each verb will have varying noun preferences, given a majority of verbs have biased noun tendencies. If the verbs were truly synonymous, theoretically, they should have interchangeably occurred with all nouns relatively equally in their distribution.

Most of the verbs such as *design*, *discover*, and *give* depict this difference very well, meanwhile some other verbs like *bring*, *obtain*, and *bestow* do not show large variations. With 17 participants, we

were able to highlight likely-dominant noun preferences, although there are some imperfections. In most cases, there is inherent association, but there are also verbs that can co-occur with more varied nouns.

One potential reason for the association is the semantic specificity of the verb. That is, there could be a semantic feature of the verb that matches the feature of the noun. For example, the strong preference of *discover* and *explore* towards *gold ore* (compared to the other nouns in the options) could be motivated by the nature of *gold ore*, which tends to be hidden and precious such that requires exploration and discovery. Another example is the unselected *gold bar* with *make* and *create* which could be due to the highly generic semantics of the verbs for the noun, which is in turn semantically more specific, especially when the *gold bar* undergoes a more specific process during production. The second potential explanation for the preference of the verb with the noun could be the entrenchment of the noun when it is frequently used with the verb in real text/corpus that participants could have heard before (see for instance, Diessel, 2016; Diessel & Hilpert, 2016; Hilpert & Diessel, 2017).

Previous works such as Yoon (2016) used a longitudinal way of researching by asking the participants to construct essays and compare them to two corpora created specifically for the research. Then, the results were collocations hand-picked from the essays. Using the Corpus of Contemporary American English (COCA), the results were then calculated in terms of frequency and Mutual Information (MI) score. Church and Hanks (1990) also incorporated MI scores to count collocations solely in a corpus instead of using real participants. Unlike these two works mentioned, the current paper simply cannot use MI score due to the fact that the verbs and nouns were pre-determined, and the sample size is too small. For the purpose of this paper, adapting a small part of their research should be sufficient.

4. CONCLUSION

With all the results above, it can be inferred that there appears an inherent association of verbs to specific nouns than others (as seen in Figure 6-Figure 10). Most of the verbs can show stark contrast in their noun preferences though there are nouns that are selected equally for a given verb (see for instance nouns for the verb *bring*). We have provided some explanation as to why certain nouns are preferred for certain verbs, given the semantics and possibly usage-frequency effect on the entrenchment of the verb-noun combination. This assumption regarding the role of frequency in explaining participants' selections of certain verb-noun combinations should be tested in future studies using collocation and quantitative techniques. Moreover, the experiment could be expanded in future studies to include more nouns and verbs involving much larger pool of participants than the one provided for this paper.

REFERENCES

Chan, T., & Liou, H.-C. (2005). Effects of Web-based Concordancing Instruction on EFL Students' Learning of Verb – Noun Collocations. *Computer Assisted Language Learning*, *18*(3), 231–251. https://doi.org/10.1080/09588220500185769

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, *16*(1), 22–29.

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, *28*(2), 130–144. https://doi.org/10.1080/09588221.2013.803982

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Diessel, H. (2016). Frequency and lexical specificity in grammar: A critical review. In H. Behrens & S. Pfänder (Eds.), *Experience Counts: Frequency Effects in Language* (pp. 209–238). De Gruyter. https://doi.org/10.1515/9783110346916-009

Diessel, H., & Hilpert, M. (2016). Frequency Effects in Grammar. In H. Diessel & M. Hilpert, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.120

Dong, J., & Lu, X. (2020). Promoting discipline-specific genre competence with corpus-based genre analysis activities. *English for Specific Purposes*, *58*, 138–154. https://doi.org/10.1016/j.esp.2020.01.005

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Hilpert, M., & Diessel, H. (2017). Entrenchment in construction grammar. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge.* (pp. 57–74). American Psychological Association. https://doi.org/10.1037/15969-004

Nebot, E. M. (2008). Corpus-based Activities in Legal Translator Training. *The Interpreter and Translator Trainer*, *2*(2), 221–252. https://doi.org/10.1080/1750399X.2008.10798775

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, *24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rajeg, G. P. W. (2020). *Tutorials for Corpus of Contemporary American English (COCA)*. 0 Bytes. https://doi.org/10.6084/M9.FIGSHARE.12362756

Ucar, S., & Yükselir, C. (2015). The Effect of Corpus-Based Activities on Verb-Noun Collocations in EFL Classes. *Turkish Online Journal of Educational Technology - TOJET*, *14*(2), 195–205.

Yamagata, S., Nakata, T., & Rogers, J. (2023). Effects of distributed practice on the acquisition of verb-noun collocations. *Studies in Second Language Acquisition*, *45*(2), 291–317. https://doi.org/10.1017/S0272263122000225

Yoon, H.-J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing*, *34*, 42–57. https://doi.org/10.1016/j.jslw.2016.11.001