P-ISSN 2355-2794
E-ISSN 2461-0275

The Use of Semantic Transparency and L1-L2 Congruency as Multi-Word Units Selection Criteria

Maryam Barghamadi[*1]
James Rogers[2]
Joanne Arciuli[1]
Amanda Müller[1]

[1]College of Nursing and Health Sciences, Flinders University, Adelaide 5042, AUSTRALIA
[2]Faculty of Foreign Studies, Meijo University, Nagoya 461-0048, JAPAN

Abstract

Multi-word units (MWUs) are defined as two or more words that commonly co-occur. There is evidence that knowledge of high-frequency MWUs is essential to language fluency, leading to growing research identifying valuable MWUs to learn and the impact of L1-L2 congruency and semantic transparency on the learning burden of MWUs. Therefore, there needs to be more research on which MWUs should be selected with these criteria. This article highlights an investigation of the role of congruency and semantic transparency using a corpus-based list that offers a sizable sample of MWUs that appear in general English. In this study, we analysed a list of 11,212 high-frequency MWUs created using a lemmatised concgramming approach to examine the role of semantic transparency and L1-L2 congruency. The list was translated into Persian, and L1-L2 congruency ratings were given to each item. The list was also classified based on Grant and Bauer's (2004) taxonomy to explore the role of semantic transparency to determine the extent to which these two factors play a role in the learning burden of the MWUs. The results showed that 85% of items were literal, and a low number of opaque items were found in the high L1-L2 congruency rating, suggesting a positive relationship between congruency and transparency. The current research then discusses the implications of these two criteria for teaching English as a

[*] Corresponding author, email: barg0008@flinders.edu.au

Citation in APA style: Barghamadi, M., Rogers, J., Arciuli, J., Müller, A. (2023). The use of semantic transparency and L1-L2 congruency as multi-word units selection criteria. *Studies in English Language and Education, 10*(2), 723-740.

Received October 21, 2022; Revised January 23, 2023; Accepted April 18, 2023; Published Online May 31, 2023

https://doi.org/10.24815/siele.v10i2.28644

second language and considers their importance in designing teaching materials.

Keywords: L1-L2 congruency, lemmatised concgramming approach, multi-word units, second language acquisition, semantic transparency.

1. INTRODUCTION

Multi-word units (MWUs) aid receptive and productive fluency (Boers, 2020) and are considered to be essential for language proficiency (Shin & Chon, 2019). Improved productive/receptive comprehension speed and more native-like fluency have become the fundamental reasons to focus on MWUs (Hoey, 2005). Therefore, a growing emphasis has been placed on MWUs in second language acquisition (SLA) research over the last two decades since a significant portion of spoken and written discourse consists of such items (Erman & Warren, 2000; Foster, 2001).

A substantial body of literature indicates MWUs are the most challenging aspect of learning English in various contexts (e.g., Laufer & Waldman, 2011; Yamashita & Jiang, 2010; Zhou, 2016). There are several reasons why learning MWUs is challenging. First, there are inconsistencies in terminology (Wolter, 2020), which creates problems for researchers and practitioners. Scholars have defined collocations and MWUs in several ways (Rogers et al., 2021, p. 143). Therefore, various terms have emerged, such as 'fixed expressions', 'formulae', 'formulaic language', 'phrasal expressions', 'prefabricated language chunks', and 'word associations.' Inconsistent use of terminology makes it challenging for teachers, students, and even novice researchers to understand the distinctions between the various types of MWUs. This study defines collocations and MWUs as one entity and uses these terms interchangeably.

Another issue is that there needs to be an agreement on what identifying criteria are used for selecting the essential MWUs for students. Much research has focused on developing resources such as general and academic collocations lists (e.g., Martinez & Schmitt, 2012; Rogers et al., 2021; Shin, 2006), but there is only a small amount of guidance on which MWUs to prioritise during instruction. For example, Martinez and Schmitt (2012) state that non-transparent or non-compositional MWUs are valuable for L2 learners. Shin (2006), on the other hand, used grammatical well-formedness as a criterion. In addition, MWUs are rarely addressed in language courses and are seldom incorporated into teaching materials and classroom activities (e.g., Boers et al., 2017).

Researchers have observed some factors that affected the learnability of L2 collocations: frequency, L1-L2 congruency, and semantic transparency. The importance of frequency has been confirmed in learning MWUs when more frequent items should be learned faster than infrequent ones (e.g., Wolter & Gyllstad, 2013). Also, much research has explored how congruent collocations (e.g., L1 translation equivalents) are less challenging than incongruent items (e.g., Yamashita & Jiang, 2010). An illustration of congruent items between Persian and English would be 'take' in 'take a photo' (عکس بگیر /ʔãks bɪgiːr/), which are word-to-word equivalents. However, 'take' in 'take medicine' is treated as an incongruent item since the verb of the phrase in Persian literally means 'eat' (بخور /bɔxɔːr/). Therefore, a word-to-word translation strategy to produce MWUs could lead to unaccepted word combinations.

Furthermore, another factor that could affect the processing of collocations is the degree of semantic transparency or how literal/figurative an MWU is. Gyllstad and Wolter's (2016) study found that participants' reaction times were shorter for more transparent word combinations. For instance, a Persian learner can more quickly understand the meanings of two words in 'take the money' that make up a literal collocation. In contrast, non-transparent combinations such as 'take a side' or 'take someone to task' could be challenging for Persian learners.

Few studies have focused on teaching L2 collocations (e.g., Laufer & Girsai, 2008). Laufer and Girsai emphasised the importance of using contrastive analysis and translation in teaching collocations based on form-focused instruction or intentional learning. By using awareness-raising activities with form-focused instructions and tools, teachers can help learners develop their collocational fluency (see Barghamadi et al., 2022). Explicit instruction regarding semantically opaque or non-transparent word combinations that frequently co-occur is recommended (e.g., Martinez & Schmitt, 2012; Moon, 1994). Since a large percentage of collocations are semantically transparent (Rogers, 2017), it may not be reasonable to focus only on non-transparent items.

Some evidence suggests that second language learners should focus on learning MWUs which are semantically non-transparent because such items have a higher learning burden than literal combinations of words (e.g., Macis & Schmitt, 2017a). Also, L1-L2 congruency should play a role in selecting items to teach when learners are prone to make errors comprehending or producing MWUs which are said differently than in their L1 (Peters, 2016).

Thus, this paper investigates whether L1-L2 congruency or semantic transparency are fundamental criteria for selecting useful English MWUs to teach explicitly to native Persian speakers in a large-scale study. This research explores the percentage of high-frequency MWUs based on L1-L2 congruency and semantic transparency by utilising Rogers' (2017) list of 11,212 high-frequency MWUs as a starting point; we addressed the following two research questions:

1. What percentage of high-frequency word combinations is semantically opaque?
2. What percentage of high-frequency English MWUs are incongruent with their Persian translations?

The current study identifies the fundamental criteria for selecting high-frequency MWUs to teach explicitly based on a modern lemmatised concgramming approach. Additionally, this study aims to provide a new perspective for second language practitioners and guidelines for incorporating MWUs into their courses. The following sections will provide an overview of the definition of MWUs and the role of L1-L2 congruency and semantic transparency.

2. LITERATURE REVIEW

2.1 Defining and Identifying MWUs

Defining MWU should start with defining what collocations are because, depending on the perspective, both share similar characteristics. For example, depending on the number of words in a phrase, Biber et al. (1999) differentiated collocations from MWUs. According to their research, collocations consist of two

words, while idioms or lexical bundles go beyond that. Depending on the approach taken, collocations are handled differently from MWU. However, frequency-based and phraseological approaches are common approaches to conducting MWU research.

Based on grammatical structure and semantic transparency, the phraseological approach (e.g., Howarth, 1998) uses a typological method to identify collocations (Gyllstad & Wolter, 2016). In this approach, word combinations with literal lexical constituents, such as 'pay a bill' or 'read a book', are considered free combinations, not MWUs. However, word combinations such as 'pay a visit', where one constituent word appears in its literal sense and the other in its figurative sense, are considered MWUs worthy of direct instruction. However, this method has a drawback because the frequency of occurrence is not utilised to identify useful collocations. Consequently, using the phraseological approach, MWUs such as 'lousy weather' could be selected for direct instruction. However, they may not be frequent enough to make them the most suitable items to study at a specific point in the learner's level of fluency.

In contrast, with the frequency-based approach (e.g., Sinclair, 1991), collocations are defined as combinations of two words which frequently occur together, regardless of how semantically transparent they may be (Macis et al., 2021). In L2 learning and teaching, collocations with low frequency are less likely to be selected using this approach. Several researchers have taken the frequency-based approach in their study; however, some of the word combinations that are classified as collocations in one study could be considered idioms in another study, such as 'pull strings' in Webb et al.'s (2013) study and 'bottom line' in Wolter and Gyllstad's (2013) study. Thus, what might be called a collocation under one approach may not be considered a collocation in another.

The frequency-based approach lets us identify valuable items using a large corpus and computation. Some search engines have been developed to identify MWUs, such as n-grams and skipgrams, in corpus data. The former recognises linear sequences. A researcher might find n-gram patterns such as 'a lot of people' but not patterns such as 'a lot of different people'. To detect non-contiguous word associations and deal with constituency variation in natural language processing (NLP), skipgrams searches have been developed (Wilks, 2005). The limitation of the skipgram approach is that it only handles three-word skipgrams and cannot handle positional variation (i.e. AB, BA). Therefore, Greaves (2005) developed the ConcGram program to cope with these limitations.

A concgram "constitutes all the permutations of the constituency and positional variation generated by the association of two or more words" (Cheng et al., 2006, p. 411). Cheng et al. (2006) showed that the concgramming approach could identify collocations with consideration for both constituency and positional variation. Constituency variation refers to one or more terms that occur between the related words (AB, ACB, e.g., 'make money', 'make some money'). Positional variation refers to related words relative to one another in various positions (AB, BA, e.g., 'world city of Asia', 'Asia's world city') (Cheng et al., 2006, p. 413). With concramming, co-occurrence is calculated by counting co-occurrence between all the inflected forms of a pivot word and a collocate with the same part of speech (a lemma). Based on this method, both constituency variation (AB, ACB) and positional variation (AB, BA) are taken into account, so structures such as 'jury's verdict,' 'jury's shocking verdict', and 'verdict of the jury' are counted together for lemma pair 'jury/verdict'.

This results in more accurate frequency counts because similar items can thus be counted together. Therefore, a concgram's associated words can account for various combination types. Some scholars believe that concgramming is ideal for identifying MWUs (Cheng et al., 2006; Durrant, 2009; Rogers et al., 2021).

In the current study, a traditional perspective of frequency of co-occurrence (Biber et al., 1999) is combined with a lemmatised concgramming approach to define and identify useful MWUs. As a result, all the MWUs identified are considered the same in this study, regardless of whether they are phrasal verbs or idioms. This perspective defines collocations and MWUs as one entity that falls between transparency and non-transparency via the lemmatised congramming approach.

2.2 L1-L2 Congruency as a Criterion for MWU Identification

Semantic constraints and grammatical rules play a role in the structure of collocations. Therefore, it is impossible to explain why some word combinations, such as 'strong coffee' or 'big mistake', are acceptable, but 'powerful coffee' or 'large mistake' are unacceptable. This aspect of collocations affects their learning burden because learners may not be aware of collocational restrictions. It has been found that collocations with a direct correspondence between L1 and L2 are easier to produce when learners can transfer knowledge from their L1 (Ellis, 2008).

Research on L2 collocational processing shows that congruent collocations with an equivalent word-to-word translation in a learner's L1 are learned faster than incongruent collocations (e.g., Gyllstad & Wolter, 2016; Nesselhauf, 2005). Due to cross-linguistic relationships, word-to-word translation will likely have a high error rate in most languages. Therefore, learners make more errors when collocations are incongruent due to L1 interference (e.g., Davoudi & Behshad, 2015; Gyllstad, 2005; Nakata, 2006; Peters, 2016; Wang & Shaw, 2008). Table 1 demonstrates examples of potential collocational errors in English based on L1-L2 congruency.

Table 1. Potential collocations errors based on L1-L2 congruency.

Example (English in parenthesis)	Language	Reference source
'look for money' (earn money) 'learn knowledge' (gain knowledge) 'bring some reasons' (state some reasons)	Persian	Davoudi & Behshad (2015)
'make a photo' (take a photo)	German	Gyllstad (2005)
'take contact' (make contact)	Japanese	Nakata (2006)
'make your homework' (do your homework) 'do a suggestion' (make a suggestion)	Dutch	Peters (2016)
'do changes' (make changes) 'do a great effort' (make a great effort) 'make damage' (do damage)	Swedish	Wang and Shaw (2008)
'have risk' (take a risk) 'have harm' (cause/do harm)	Chinese	Zhou (2016)

All of the studies mentioned in Table 1 concluded that learning incongruent collocations is more complicated than learning congruent collocations. Although there is still much to learn about the factors influencing L2 collocational processing, there is agreement that congruency is a fundamental factor to consider. The role of L1 interference and its persistent effect on L2 collocation acquisition has made it an important criterion to consider in collocational research when researchers endeavour

to explore the relationship between congruency and other factors such as level of proficiency (e.g., Özdem-Ertürk, 2021; Sonbul & El- Dakhs, 2020) and such studies have concluded that the best predictor of collocational accuracy and ease of learning is congruency with the learner's L1.

Despite its importance, this criterion has yet to be addressed in developing teaching materials. To illustrate, Shin (2006) points out that L1-L2 congruency is a crucial factor; however, he compared only 10% of the English collocations in his study with Korean. To date, Rogers' (2017) study is the only one examining L1-L2 congruency on a large scale (11,212 English MWUs). He found that approximately half of the items were incongruent with their Japanese counterparts.

Providing a collocation resource for learners to help them avoid errors due to L1 influence would be very useful. However, our search yielded no studies that elicited useful MWUs based on congruency for Persian learners. Considering these findings from a teaching and learning standpoint, it appears useful for teachers to highlight differences between L2 words and their L1 translations so that learners can avoid producing strange word combinations and more accurately comprehend incongruent combinations. Therefore, focusing on incongruent collocations is essential for L2 learners. Still, it does not negate the importance of learning congruent items because they also comprise many high-frequency MWUs.

2.3 Semantic Transparency as a Criterion for MWU Identification

Collocations can be classified based on semantic transparency. This is referred to as the phraseological view of collocation. This view purports that collocations consist of at least one word that is not semantically transparent (Wolter & Yamashita, 2015). For example, 'pay the bill' is a free combination due to the literal meanings of both words, but 'pay attention' and 'pay a visit' are collocations since 'pay' is not literal in these combinations.

The stance that free combinations include word combinations in which each word takes its literal meaning, and figurative combinations consist of words taking a non-literal meaning has been used in several studies which aim to classify collocations (Cowie, 1988, 1994, 2001; Grant & Bauer, 2004; Howarth, 1998). In general, all classifying criteria for collocations have come from studies based on the phraseological approach. Specifically, Grant and Bauer (2004) divided MWUs into four categories:

1. Literal/Compositional: The meaning of MWUs is transparent or closely related to each item (e.g., 'hit the ball', 'break eggs')
2. Once: When one word of an MWU is non-literal or non-compositional (e.g., 'driven to quit')
3. Figurative: Structures such as 'hit the nail on the head' and 'give someone the green light' are not literal, but such combinations can be "reinterpreted pragmatically" to be comprehended (Grant & Bauer, 2004, p. 51)
4. Core idiom: The meanings of all the individual words are entirely unrelated to the meaning of the idiom as a whole (e.g., 'by and large')

Since semantic transparency could be another criterion for identifying MWUs, some researchers agree that it is helpful to classify collocations into literal, figurative, and core idioms in language learning (Grant & Nation, 2006; Nation, 2020). Recently, Macis and Schmitt (2017a) classified 54 collocations into literal (78%), figurative

(3.7%), and duplex collocations (18.5%). Since duplex collocations refer to collocations with both literal and figurative meanings, they concluded that it is essential to consider figurative meanings when teaching such collocations. Based on semantic transparency, Yamashita (2018) categorised 240 collocations employed in five experimental studies as congruent and incongruent. His research confirmed that transparent items dominate the congruent category and opaque items dominate the incongruent category.

More transparent meanings seem to be learned before those less transparent (Macis & Schmitt, 2017b, p. 324). For example, Persian learners are likely to understand and more easily learn the semi-transparent MWUs ('take care') compared to more opaque items ('take someone to task'). Similarly, scholars such as Moon (1994, 1997) and Van der Meer (1998) believe that only entirely semantic non-transparent word combinations should be considered collocations. With such an approach, most high-frequency collocations would be ignored if collocations consist of one non-transparent word.

In addition to the issue of L1-L2 congruency, there needs to be more agreement in the research regarding teaching literal or non-literal collocations. If a high proportion of high-frequency combinations are semantically transparent, then there would be an argument for their consideration for direct instruction. Therefore, this research investigates the percentage of high-frequency MWUs into semantic transparency categories using Grant and Bauer's (2004) taxonomy and gives each MWU L1-L2 congruency rating after the list is translated into Persian.

3. METHOD

3.1 Research Design

The present research is a partial replication study of Rogers' (2017) study while replacing the comparison between Japanese and Persian. It is also a follow-up study of Barghamadi et al. (2023), which emphasised the role of L1-L2 congruency between English and Persian but ignored semantic transparency. Thus, the current study investigates transparency and congruency's role in identifying useful MWUs for direct teaching.

3.2 Data Source

Here we define collocations and MWUs as one entity. Collocations (a pivot word and a collocate) are presented to learners within the chunks of the language they most commonly occur together with. Selecting a well-constructed corpus is essential to creating such a resource. Researchers such as Dang (2020) agree that the nature of corpora impacts the result of any word list derived from its data, and thus it should be selected carefully. The British National Corpus (BNC) and Corpus of Contemporary American English (COCA), which consist of 100 million tokens and 450 million tokens, respectively, have been utilised in many studies to develop word lists (e.g., Durrant, 2014) and are recommended by many scholars (e.g., Dang, 2017; Dang & Webb, 2016; Rogers et al., 2021). Both are suitable for English L2 learners because of lexical coverage, students' vocabulary knowledge, and educator assessments of word

usefulness (Dang & Webb, 2016). Although the COCA only contains American English, it is not only much bigger than the BNC, but it is also significantly more updated. The COCA also provides a good balance between spoken and written materials from spoken, academic texts, newspapers, popular magazines, fiction, TV and movie subtitles, blogs, and web page genres. In Rogers' (2017) study, the COCA was selected for the following reasons:

- Development of the BNC largely ceased in 1993
- The COCA is four times larger than the BNC
- The researcher is an American native speaker and aimed to create an English American resource.

It is worth noting that the BNC has been updated partly with the British National Corpus 2014 (Love et al., 2017) with the addition of a 100 million-word "real life" spoken English section. However, this improvement to the 100 million words of the mostly formal language of the original BNC still leaves it being less than half the size of COCA. The COCA's balance and variety of its sections still give it an advantage as well.

In his study, Rogers (2017) aimed to identify the most exemplary MWUs for high-frequency lemmatised concgrams for general English considering various factors, including frequency, mutual information, semantic transparency, L1-L2 congruency, dispersion, and chronological stability. Concerning the lemmatised concgramming method, he used co-occurrence of lemma such as 'take' and 'break' as 'take a break', 'take breaks', 'taking breaks', 'took a break', and 'take a quick break' were counted as one unit. It is possible to identify 'take a break' as the most frequent lemma constituent within concordance strings that contain both 'take' and 'break'. The frequency counts in an MWU list would be skewed by counting these MWUs separately rather than with the concgramming method.

Also, separating such items may result in redundant teaching if counted separately. Rogers' (2017) study demonstrated the advantages of this methodology over previous methods of identifying collocations/MWUs, so the current research has used Rogers' methodological approach to investigate Persian-speaking learners. This study used his list because it was the most extensive known list created for second language learners based on the most modern method of MWU identification: lemmatised concgramming.

3.3 Instrument

Rogers' (2017) list of 11,212 high-frequency MWUs was used as the primary data source in this study because it was the largest known list created for second language learners based on the most modern method of MWU identification: lemmatised concgramming. Barghamadi et al. (2023) translated Rogers' list into Persian, and each MWU was given an L1-L2 congruency rating of 0-12 (with 12 denoting total congruency). A 12-point scale was used because most of the MWUs were three to four words long, and thus the rating system was simplified by having an even number of points per word for the vast majority of items. For example, when MWUs consist of 3 words and have the same meaning in both Persian and English, each word will receive 4 points, which equates to 12 points in total.

Give an example, 'do your homework' consists of 3 words and is a word-for-word equivalent of its Persian translation (تکلیف خود را انجام بده/takali:f xo:d ra: anʒa:m

bɪdɪh/). In comparison, 'make a mistake' (اشتباه کردن/ɪstɪba:h kardan/) received 6 points since 'make' does not have the same individual word meaning in Persian. Since Barghamadi and her colleagues did not use multiple raters, this list rerates by utilising the12 points system and compares with semantic transparency classification simultaneously.

For semantic transparency ratings, two raters used Grant and Bauer's (2004) taxonomy to determine whether or not the MWUs were literal, once, figurative, or core idiom. This current study followed Rogers' (2017) study using Grant and Bauer's taxonomy. However, he added one classification he called 'outliers' for items that fell outside the taxonomy, such as MWUs with polysemy issues (e.g., 'bear children'). To classify MWUs, the following protocol was used:

- **Literal (12 points):** Each content word (noun, verb, adjective, adverb) literally means what it means (e.g., 'black cat', 'go to work').
- **Once (8 points):** One or more content words are figurative, and at least one word is literal such as 'long face' ('face' is literal, but 'long' is figurative) and 'dog days '('days' is literal, but 'dog' is figurative).
- **Outlier (6 points):** Items contain a homonym (polysemy issue) that can be easily misunderstood (the significantly rarer homonym is used, such as 'bear children', or situations where the meaning is particular, such as 'intensive care', 'social security', or 'foster care'. Ratings of 6 are also reserved for items in which language is used colourfully, such as 'in harm's way.' Also, 6 points were given to items where a preposition is used in a way that is very different from its literal meaning, such as 'I sort of think' (in this case, 'of' is meaningless to some extent). In addition, if an MWU seems to be formed arbitrarily (there is no rhyme/reason why a particular word is used and not another more logical one), it was also given 6 points. For instance, why 'record label' and not 'record company'? Why 'cast a shadow' and not 'put a shadow'?
- **Figurative (4 points):** The entire phrase is figurative, but the meaning can be inferred. Most people think of these as idioms, but they are not. True idioms cannot be understood by their parts and are called 'core idioms.' So, figurative words are what most people think idioms are (e.g., 'as good as gold', 'hit the nail on the head').
- **Core Idiom (0 points):** The entire phrase is figurative, and the meaning cannot be inferred from the individual words (e.g., 'shoot the breeze', 'a piece of cake'). Other examples of items that received 0 points include 'got under way' (it means something has started but has a high potential for confusion. Nothing is 'gotten', nothing is going 'under' anything, and 'way' is usually used for direction or a method of doing something, of which the typical usage of this phrase has nothing to do with) and 'wind up' (nothing is being 'wound' and nothing is going 'up'. Thus, it would be difficult for a learner to guess the meaning).

When there could be multiple interpretations of a phrase, one often was literal, and the other was figurative. The raters compared the more common usage of the MWU with the Persian translation. However, when it still needed clarification, the raters looked at the use in the example sentence and rated accordingly. Inter-rater reliability was then conducted to confirm the reliability of the semantic transparency classifications and the L1-L2 congruency ratings.

4. RESULTS

The subjective nature of assigning semantic transparency ratings can create reliability and replicability issues. However, in this current study, interrater reliability was 97%. Researchers have deemed interrater reliability to be from 75% to 90% (Larson-Hall, 2015; Stemler, 2004), and thus the level of agreement found in this study points to the protocol being reliable. Two raters marked some items differently and were not literal (mostly, these were figurative or core idiom). For example, 'give him the benefit of the doubt' may be classified as a core idiom and figurative by different raters. These occasional minor difference between raters was not crucial to the main focus and results of the study since the main focus is the literal interpretation of MWUs. In such cases, the results of the rater, as a native speaker and English university lecturer, were calculated in the analysis. The results of this part of the study can be seen in Table 2 with the addition of L1-L2 rating categorisation.

Table 2. Semantic transparency of the MWUs with L1-L2 rating categorisation.

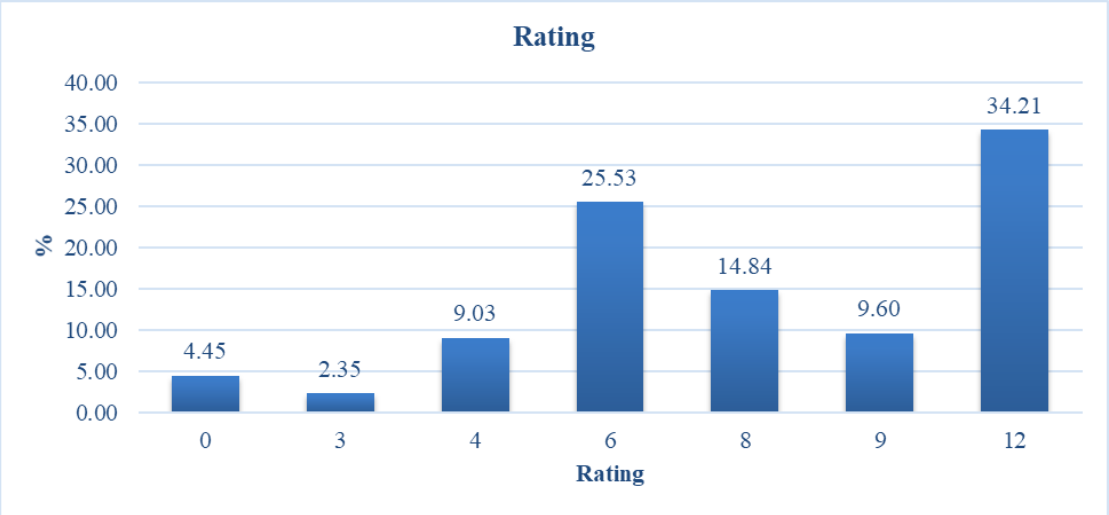
L1-L2 Rating	Literal (12)	Once (8)	Outlier (6)	Figurative (4)	Core Idiom (0)
0-3	329	142	82	83	124
4-6	3,115	343	295	75	47
8-9	2,513	114	80	26	8
12	3,656	78	67	13	3
Total	85.9%	6.04%	4.68%	1.75%	1.63%

As relatively few items were judged to be once, figurative, core idiom, or outlier, these categories were combined into one opaque category, and 14.1% fell within it. This result made it salient that many of the MWUs examined were classified into the literal category.

The results of the L1-L2 rating can be seen in Figure 1. Out of 11,212 MWUs, 7,376 (65.78%) received a rating of 0-9. A total of 3,836 MWUs (34.21%) received a rating of 12. Thus, nearly two-thirds of the items examined may pose a higher learning burden because of incongruences with their translations. If we consider the incongruent items when half of the words are not equal to their translation equivalent (a rating cut-off of 6 or less), 41.3% of items fall under this category. Consequently, it became clear that many of the MWUs examined were not congruent with Persian. According to Table 2, there is a positive relationship between low L1-L2 ratings and opaque semantic transparency. As L1-L2 congruency ratings increased, the number of items classified as figurative and core idiom items decreased. Table 3 provides examples of semantic transparency classification and L1-L2 congruency rating between Persian and English.

Table 3. Sample of MWUs with semantic transparency classification and congruency with Persian.

MWUs	L1-L2 rating	Semantic transparency
eye to eye	0	0
made up his mind	3	0
caught my eye	4	4
given name	6	6
took a deep breath	8	8
make it difficult for	9	12
leave me alone	0	12



Note. A rating of 12 denotes total congruency.

Figure 1. L1-L2 (Persian/English) congruency ratings of high-frequency English MWUs.

It is worth mentioning that in the first step, native English speakers classified the semantic transparency of items. Next, one researcher ran the L1-L2 congruency rating that led to finding MWUs with various meanings. The rater found 1% of items with literal and opaque meanings that Macis and Schmitt (2017a) call duplex collocations, such as 'a piece of cake,' 'the bottom line,' and 'inner circle'. For example, the phrase 'a piece of cake' has one literal meaning and one figurative meaning ('as easy as eating a piece of cake'). These could be signals in the teaching process that some items have two meanings to consider.

We concluded from our data set that most collocations follow a traditional word combination where two or more literal meanings can be added (A +B+ and so on). In contrast, our analyses revealed that most of these items are somewhat incongruent with Persian. For example, 'make sacrifices' and 'make a decision' have literal meanings but are incongruent with Persian when the verb is equal to 'do' and 'take', respectively. Therefore, only focusing on traditional patterns of word combinations could lead to unacceptable structures.

On the other hand, a small proportion of our sample collocations (1,518 items or 14.1%) have opaque meanings. The evaluation of these items indicates that only 161 items are opaque but congruent with Persian. For example, 'my heart stopped' is classified as figurative when the meaning refers to 'it gives you a sudden intense feeling of fear,' still, it is a congruent word for word with its Persian translation and used in the same fashion. Therefore, the combination of both opaque items and duplex items makes up 15.1% of total items. In comparison, a substantial percentage of the items in the present study are incongruent even when using a cut-off of 6 points (41%).

5. DISCUSSION

Knowing a word includes knowing its semantics and collocates (Nation, 2013). The literature indicates that MWUs make up a significant amount of spoken and written discourse (e.g., Erman & Warren, 2000), and knowledge of these could

facilitate L2 learning, help learners become more fluent, and solidify their status as expert language users (Siyanova-Chanturia & Pellicer-Sánchez, 2019).

In accordance with the lexical approach to MWUs, it is recommended that learners observe recurring lexical chunks in L2 input (Lewis, 1993). This perspective has been supported by researchers such as Pellicer-Sánchez (2020), who stated that MWUs "should be a component of the vocabulary learning curriculum" (p. 158). Therefore, knowing what words co-occur is crucial to developing language fluency. However, several studies found that MWUs were excluded from teaching materials (e.g., Boers et al., 2017), and there seems to be less agreement regarding inclusion and inclusion criteria.

The selection criteria value can also be confirmed by determining the percentage of the items semantically transparent. In the current study, we found that literal collocations made up the vast majority of the items examined. This result is in line with previous studies (e.g., Rogers, 2017; Macis & Schmitt, 2017a), which found that most of the items examined were also literal. For such items, it would seem that learners do not need to consider the meanings of the collocations since they need to understand the meanings of the component words. Although when understanding the meaning of a particular phrase is a concern, knowing which words arbitrarily go together can be a challenge, i.e., to recognise why 'strong' comes frequently occurs with 'coffee' and not 'powerful'. In this regard, Shin and Chon (2019) claim that L2 learners have difficulties with all types of MWUs with varying levels of transparency, and thus, such items also deserve study time.

There is no question that high-frequency items are valuable; however, the learning burden of these items may be influenced by other factors, such as L1-L2 congruency. The value of selecting criteria can be confirmed by determining the extent to which high-frequency English MWUs are congruent with the target language. The current study fills the gap in the literature for Persian-speaking learners. The findings also support previous research showing that L1 interference is the main reason for L2 learners' errors (e.g., Davoudi & Behshad, 2015). According to the results of this study, a large proportion of the MWUs examined were incongruent with Persian to some extent (e.g., 'do a mistake' instead of 'make a mistake').

Also, the present study's findings revealed a positive relationship between L1-L2 ratings and semantic opacity. This finding agrees with Yamashita's (2018) research. Albeit a small-scale study, Yamashita found that transparent items dominated the congruent category. In contrast, opaque items dominated the incongruent category, and he concluded that semantics may have played a role in congruency effectiveness. Therefore, it would be helpful to determine the best way to teach learners these items. Macis and Schmitt (2017a) pointed out that most collocation pedagogy and textbooks have centred on word combinations with literal meanings. For example, 'strong coffee' not 'powerful coffee' or 'big mistakes' not 'large mistakes' may emphasise in English courses and understanding how words co-occur.

Since collocations are identified and defined differently using different analytical methods, it is unsurprising that there is a need for comprehensive resources. Also, teachers and materials developers need more guidance in selecting items that might be particularly noteworthy for students. Therefore, there still needs to be more research and materials regarding the semantic opacity of MWUs. It would be desirable if teachers had access to textbooks and other resources that used congruency to select essential items for teaching and identify the most difficult MWUs.

To help teachers determine what phrases to teach, Martinez (2013) suggested using the Frequency Transparency Framework (FTF) approach. Using this, a frequent and opaque phrase would be introduced first, followed by a frequent and transparent phrase, then a less frequent but opaque phrase, and finally, the least frequent and more transparent expression. Martinez notes that the framework can only be practical when frequency measures are adjusted to reflect learners' needs. For example, it is unlikely that any English language teacher or writer would insist that 'take the bus' should be taught late in any language course; while this MWU is not very common in the BNC (Martinez, 2013, p. 192), it is a critical element in giving directions. On the other hand, if the most challenging aspect of acquiring L2 knowledge is related to opaque items, then concentrating on MWUs with non-literal meanings is a valuable and essential teaching strategy.

Nevertheless, if many collocations are literal and L1-L2 congruency is the primary source of L2 errors, extra attention should be given to this cross-linguistic issue. In the current research, literal formulations were the most numerous high-frequency collocations, even though meaningfully opaque collocations deserve more attention than transparent ones, suggesting that a total focus on non-literal items is unreasonable. Nonetheless, many L2 errors are rooted in L1 interference (e.g., Davoudi & Behshad, 2015; Peters, 2016), and knowledge of collocation can be predicted by congruency (Nguyen & Webb, 2017). Therefore, items with low L1-L2 congruency deserve more consideration in language pedagogy.

Overall, all types of MWUs with non-transparent meanings make up a small sample of items examined in this current study. While most items are transparent, understanding their meanings depends on the learner's prior knowledge. For example, to interpret the item 'take a break here' and 'take medicine' (which are classified as literal), it is essential to know that the first 'take' is equal to 'have' and the second is equal to 'eat'. So, despite being classified as literals, they are incongruent with their Persian equivalent and, therefore, can be challenging to comprehend. These examples demonstrate why focusing on L1-L2 congruency is an essential criterion to consider.

From a teaching perspective, a widespread endorsement is given to raising learners' awareness of collocations as integral to language and encouraging them to pay attention to collocations. Due to the opaque nature of collocations, teachers must be aware that they are difficult for students to comprehend. Current literature indicates that an explicit approach to teaching collocations would be ideal. Thus, further research is needed to help determine how to identify such items and how to teach them effectively.

6. CONCLUSION

MWUs are widely considered essential for language acquisition. High-frequency items are unquestionably beneficial, but other criteria, such as L1-L2 congruency and semantic transparency, may affect their learning difficulty. This study revealed that a high ratio of the MWUs examined was classified as literal formulations and incongruent with Persian. Therefore, studies that only view literal formulations as collocations will end up ignoring a large proportion of the high-frequency MWUs this study examined. This is problematic since their incongruence with Persian means there is a high chance of error and a need for them to be taught. In addition, a low number

of opaque items were assigned high L1-L2 congruency ratings in this study. L1-L2 congruency is critical in the teaching process compared to emphasising semantic transparency only. However, despite this study having very high interrater reliability, the number of raters and the unavoidable subjective nature of the measures examined create some limitations in how the results of this study can be interpreted. Thus, more research is called for in this regard.

Additionally, to the best of our knowledge, there were few studies to compare with our findings due to our study's novel approach and scale. Therefore, future research should attempt to address any shortcomings that exist. However, despite these limitations, we feel that this study provides an excellent first step toward creating a valuable resource for Persian-speaking learners of English to use to improve their MWU fluency.

REFERENCES

Barghamadi, M., Rogers, J., Arciuli1, J., Han, W., & Muller, A. (2023). L1-L2 congruency as a criterion to identify collocations based on contrastive analysis. *Australian Journal of Applied Linguistics, 6*(1), 1-14. https://doi.org/10.29140/ajal.v6n1.716

Barghamadi, M., Rogers, J., & Muller, A. (2022). On the learning of multi-word units via flashcard applications. *Australian Journal of Applied Linguistics*, *5*(1), 1-18. https://doi.org/10.29140/ajal.v5n1.643

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education.

Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 143–157). Routledge. https://doi.org/10.4324/9780429291586-10

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research, 21*(3), 362–380. https://doi.org/10.1177/1362168816651464

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411–433. https://doi.org/10.1075/ijcl.11.4.04che

Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. J. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 126–139). Longman.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.). *The encyclopedia of language and linguistics* (pp. 3168–3171). Oxford University Press.

Cowie, A. P. (2001). Speech formulae in English: Problems of analysis and dictionary treatment. *Groninger Arbeiten zur germanistischen Linguistik, 44*, 1-12.

Dang, T. N. Y. (2017). *Investigating vocabulary in academic spoken English: Corpora, teachers, and learners* [Doctoral dissertation, Victoria University of Wellington]. Open Access Te Herenga Waka-Victoria University of Wellington. https://doi.org/10.26686/wgtn.17060051.v1

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 299–303). Routledge.

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics*, 167(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Davoudi, M., & Behshad, A. (2015). Collocational use: A contrastive analysis of strategies used by Iranian EFL learners. *Theory and Practice in Language Studies*, *5*(12), 2646–2652. https://doi.org/10.17507/tpls.0512.29

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157–169. https://doi.org/10.1016/j.esp.2009.02.002

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics, 35*(3), 328–356. https://doi.org/10.1093/applin/amt016

Ellis, R. (2008). *The study of second language acquisition* (2nd ed). Oxford University Press.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning. Teaching and testing* (pp. 75–93). Longman.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics 25*(1), 38–61. https://doi.org/10.1093/applin/25.1.38

Grant, L., & Nation, P. (2006). How many idioms are there in English? *International Journal of Applied Linguistics, 151*(1), 1–14. https://doi.org/10.2143/itl.151.0.2015219

Greaves, C. (2005, June, 25-29). *Introduction to ConcGram*© [Paper presentation]. Tuscan Word Centre International Workshop, Certosa di Pontignano, Tuscany, Italy.

Gyllstad, H. (2005). Words that go together well: Developing test formats for measuring learner knowledge of English collocations. *International Journal of English Studies 7*(2), 127–157.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, *66*(2), 296–323. https://doi.org/10.1111/lang.12143

Hoey, M. (2005). *Lexical priming: A new theory of words and language* (1st ed.). Routledge. https://doi.org/10.4324/9780203327630

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44. https://doi.org/10.1093/applin/19.1.24

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716. https://doi.org/10.1093/applin/amn018

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Language Teaching Publications.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

Macis, M., & Schmitt, N. (2017a). The figurative and polysemous nature of collocations and their place in ELT. *ELT Journal, 71*, 50–59. https://doi.org/10.1093/elt/ccw044

Macis, M., & Schmitt, N. (2017b). Not just 'small potatoes': Knowledge of the idiomatic meanings of collocations. *Language Teaching Research*, *21*(3), 321–340. https://doi.org/10.1177/1362168816645957

Macis, M., Sonbul, S., & Alharbi, R. (2021). The effect of spacing on incidental and deliberate learning of L2 collocations. *System*, *103*, 102649. https://doi.org/10.1016/j.system.2021.102649

Martinez, R. (2013). A framework for the inclusion of multi-word expressions in ELT. *ELT Journal 67*(2), 184–198. https://doi.org/10.1093/elt/ccs100

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Moon, R. (1994). The analysis of fixed expressions in text. In M. Coulthard (Ed.), *Advances in written text analysis,* (pp. 117–135). Routledge.

Moon, R. (1997). Vocabulary connections: multi-words items in English. In N. Schmitt & M. McCarthy (Eds.). *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). Cambridge University Press.

Nakata, T. (2006). English collocation learning through meaning-focused and form-focused activities: Interactions of activity types and L1-L2 congruence. In M. Nakano & Park K.-J. (Eds.), *Proceedings of the 11th Conference of Pan-Pacific Association of Applied Linguistics* (pp. 154–168). PAAL.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 15–29). Routledge.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Publishing Company.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research, 21*(3), 298–320. https://doi.org/10.1177/1362168816639619

Özdem-Ertürk, Z. (2021). *Factors affecting receptive and productive knowledge of collocations of tertiary level learners of English in Turkey* [Doctoral dissertation, University of Hacettepe University]. Hacettepe University Institutional Repository. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/25277

Pellicer-Sánchez, A. (2020). Learning single words vs. multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary students* (pp. 158–173). Routledge.

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113–138. https://doi.org/10.1177/1362168814568131

Rogers, J. (2017). *What are the collocational exemplars of high-frequency English vocabulary? On identifying MWUs most representative of high-frequency lemmatized concgrams* [Doctoral dissertation, University of Southern Queensland]. University of Southern Queensland Repository. https://doi.org/10.26192/5bf5ff14ed350

Rogers, J., Müller, A., Daulton, F. E., Dickinson, P., Florescu, C., Reid, G., & Stoeckel, T. (2021). The creation and application of a large-scale corpus-based academic multi-word unit list. *English for Specific Purposes*, *62*, 142–157. https://doi.org/10.1016/j.esp.2021.01.001

Shin, D. (2006). *A collocation inventory for beginners* [Doctoral dissertation, Victoria University of Wellington]. Open Access Te Herenga Waka-Victoria University of Wellington. https://doi.org/10.26686/wgtn.16945729.v1

Shin, D., & Chon, Y. V. (2019). A multiword unit analysis: COCA multiword unit list 20 and collogram. *Journal of Asia TEFL*, *16*(2), 608–623. https://doi.org/10.18823/asiatefl.2019.16.2.11.608

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). (2019). Formulaic language setting the scene. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 1–16). Routledge.

Sonbul, S., & El-Dakhs, D. (2020). Timed versus untimed recognition of L2 collocations: Does estimated proficiency modulate congruency effects? *Applied Psycholinguistics*, *41*(5), 1197–1222. https://doi.org/10.1017/S014271642000051X

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*, Article 4. https://doi.org/10.7275/96jp-xz07

Van der Meer, G. (1998). Collocations as one particular type of conventional word combinations: Their definition and character. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin, & S. Theissen (Eds.), *Proceedings of the 8th Euralex Conference* (pp. 313-322). Euralex.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME journal*, *32*, 201–232.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Wilks, Y. (2005, June 31 – July 5). *REVEAL: The notion of anomalous texts in a very large corpus*. Tuscan Word Centre International Workshop, Certosa di Pontignano, Tuscany, Italy.

Wolter, B. (2020). Key issues in teaching multiword items. In S. Webb. (Ed.), *The Routledge handbook of vocabulary studies* (pp. 493-510). Routledge. https://doi.org/10.4324/9780429291586-31

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second*

Language Acquisition, *35*(3), 451–482. https://doi.org/10.1017/S0272263113000107

Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*(5), 1193–1221. https://doi.org/10.1017/S0142716414000113

Yamashita, J. (2018). Possibility of semantic involvement in the L1-L2 congruency effect in the processing of L2 collocations. *Journal of Second Language Studies*, *1*(1), 60–78. https://doi.org/10.1075/jsls.17024.yam

Yamashita, J., & Jiang, N. (2010). L1 Influence on the acquisition of L2 collocations: Japanese ESL Users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4),647–668. https://doi.org/10.5054/tq.2010.235998

Zhou, X. (2016). A corpus-based study on high frequency verb collocations in the case of "HAVE". *International Forum of Teaching and Studies*, *12*(1), 42–50.