



Castledown

 OPEN ACCESS

Australian Journal of Applied Linguistics

ISSN 2209-0959

<https://www.castledown.com/journals/ajal/>

Australian Journal of Applied Linguistics, 6(1), 1–14 (2023)
<https://doi.org/10.29140/ajal.v6n1.716>

L1-L2 congruency as a criterion to identify collocations based on contrastive analysis



MARYAM BARGHAMADI ^a 

JAMES ROGERS ^b 

JOANNE ARCIULI ^c 

WEIFENG HAN ^d 

AMANDA MULLER ^e 

^a *Flinders University, AUSTRALIA*
barg0008@flinders.edu.au

^b *Meijo University, JAPAN*
jrogers@meijo-u.ac.jp

^c *Flinders University, AUSTRALIA*
joanne.arciuli@flinders.edu.au

^d *Flinders University, AUSTRALIA*
weifeng.han@flinders.edu.au

^e *Flinders University, AUSTRALIA*
amanda.muller@flinders.edu.au

Abstract

In second language learning research, L1-L2 congruency refers to the similarity in form and meaning of the way a word or phrase is said between two languages. If there is low L1-L2 congruency, the learning burden of a word or phrase can be higher, and thus, such items should be focused on more in teaching to help learners avoid errors. For example, an unacceptable structure could be produced when second language learners use direct or literal translation to formulate multiword units (MWUs) in the target language. This study used L1-L2 congruency as a criterion to identify English MWUs with low congruency with their Persian equivalents to create a resource for Persian-speaking English learners. An L1-L2 contrastive analysis was conducted on a list of approximately 11,000 high-frequency MWUs. This list was translated, and the items were given ratings regarding their congruency in Persian. The results revealed that more than 65% of the items examined were incongruent with their word-for-word translations, demonstrating that L1-L2 congruency is a significant factor in selecting MWUs for Persian-speaking English learners. Since focusing on incongruent items may help learners avoid errors, the results of this study can help improve the efficacy of English acquisition by Persian-speaking learners through its identification of particular MWUs that have low congruency between the two languages.

Keywords: congruency, L1 influence, multiword units, collocation, contrastive analysis

Copyright: © 2023 Maryam Barghamadi, James Rogers, Joanne Arciuli, Weifeng Han, Amanda Muller. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are in this paper.

Introduction

The definition of collocations or MWUs has been controversial among researchers. There have been a variety of terms used throughout the research process, such as *expressions*, *fixed combinations*, *formula units*, *formulas*, *lexical phrases*, *multiword lexical units* (MLUs), *phrases*, *prefabricated*, *ready-made utterances*, *recurrent combinations*, *word-like units*, and so on (Kjellmer, 1994, p. xi). MWUs, multiword items (MWIs), and multiword expressions (MWEs) are further umbrella terms that capture this endless space (Masini, 2019). There has been growing interest in collocation and MWU research in second language acquisition studies. This is because of the significant role collocations/MWUs play in providing a high level of communicative fluency for language learners in spoken and written discourse.

Collocations are the written and spoken discourse building blocks that help learners achieve receptive and productive fluency (see Boers, 2020) and efficient comprehension. Researchers agree that collocations are retrieved from memory as whole units (Boers, 2020; Nation, 2001, 2013, 2020; Schmitt & Carter, 2004; Wray, 2002), and collocational knowledge is recognised as a crucial factor for achieving fluency (Durrant & Schmitt, 2009; Pawley & Syder, 1983; Williams, 2002). For instance, valuable expressions such as *happy birthday*, *I'd like* and *on the other hand*, are retrieved from memory as whole units, leading to productive fluency (Boers, 2020, p. 143). The holistic form of MWUs helps learners improve their fluency (Nation, 2020) and provides efficient language processing (Nation, 2001; Snellings et al., 2002). Consequently, efficient comprehension occurs when “the recipient can understand the meaning of a passage of text without having to attend to every word” (Hunston & Francis, 2000, p. 270). As a result, being equipped with this phrasal processing knowledge could be considered fundamental for English learners to achieve a high level of accuracy and fluency.

One way to analyse the learning burden of L2 collocations is to focus on whether or not their meaning is congruent with their L1 counterpart. Congruent collocations have word-by-word equivalents in L1 and L2, while incongruent collocations do not have identical translations in L1, whether fully or to some extent. For example, *do homework* is equal to Persian word to word, but *make* in *make a mistake* is not equal. Researchers agree that where there is the same form and meaning between L1 and L2 in both languages, using the L1 can lead to quicker language acquisition (Ellis, 2008), while negative transfer, or interference, occurs when there are differences between them (Thao, 2020).

Congruency as a factor in second language acquisition has become a prominent topic in recent collocational research (e.g., Du et al., 2021; Özdem-Ertürk, 2021). By focusing on similarity and dissimilarity, L1-L2 contrastive analysis may be a fundamental criterion to identify collocations that need particular focus to help learners avoid errors due to L1 influence. However, this criterion seems to have been ignored in developing materials, and instead, generic textbooks have been written. The L1-L2 contrastive analysis is likely too complex and time-consuming. However, there are cases where this has occurred. For example, Rogers (2017a) innovated a lemmatised conprogramming approach, identified over 11,200 MWUs and developed a new resource for Japanese learners considering L1-L2 congruency. Thus, this paper aims to expand upon Rogers' research by using L1-L2 congruency as a criterion to select useful English MWUs for Persian learners.

Theoretical Background and Literature Review

Definition of collocations

The diversity of collocation definitions indicates the complexity of the phenomenon. According to Nation (2020, p.21), defining a collocation, and applying its definition consistently, are challenging aspects of collocation research. The frequency-based and phraseological approaches are two dominant approaches in collocational research that attempt to define this concept. Wolter and Yamashita (2015) note that to define “collocation” from a phraseological standpoint, at least one word must be semantically non-transparent. For example, *reading a book* is a free combination, but *read one's mind* is a collocation. Therefore, the phraseological approach only recognises the semantic relationship between constituents (Henriksen, 2013) of MWUs instead of considering the frequency of word co-occurrence. Instead, the frequency-based definition of collocation refers to statistically significant

co-occurrence, regardless of whether or not there are any particular semantic relations between collocating elements (Moon, 1998).

From the standpoint of frequency not being a sufficient criterion, some scholars add that being grammatically well-formed as a criterion to define and identify collocations (e.g., Kjellmer, 1987; Shin, 2006). Using grammatical structure and frequency as criteria to identify MWUs, word combinations such as *make a* or *run into* are not identified; in contrast, *make a decision* or *run into problems* are. Some scholars even further argue that to define collocations; there needs to be a distinction between collocations and phrasal verbs, prefabricated patterns, and idioms (e.g., Behns, 1993; Benson et al., 1986; Fellbaum, 2015; Hill, 2000; Woolard, 2000).

For instance, Benson et al. (1986) use the noun *murder* to demonstrate the distinguishing feature of free combination, idioms, and collocations. As a free combination, the noun *murder* can follow several verbs, e.g., *to analyse, boast of, condemn, discuss*, etc. On the other hand, idioms are frozen terms with meanings that do not correspond to the meanings of their constituent parts, for example, *to scream bloody murder*. Compared to idioms, fixed combinations or collocations (*commit murder*) fall somewhere between idioms and free combinations, and their meanings represent the sense of their constituent parts (Benson et al., 1986). Fellbaum (2015) states that idioms, like collocations, differ in how lexically and syntactically fixed they are. In comparison to collocations, which are compositional, idioms are semantically opaque to varying degrees. In contrast, Hill (2000) notes that collocations are idiomatic and predictable word combinations, and both idioms and phrasal verbs are collocations. Cruse (1986) believes a continuum exists between collocations and idioms.

In some studies, such as Rogers (2017a), collocations and MWUs are defined as one phenomenon within the lemmatised concgram method. Concgramming (Cheng, Greaves, & Warren, 2006) counts co-occurrence by counting all the inflected forms of pivot words with the same part of speech (a lemma) with a frequently co-occurring collocate (also in lemma form). This method can account for both constituency variation (AB, ACB) and positional variation (AB, BA) and therefore counts structures such as *this study found, studies find, and finding a study* together for the lemma pair *study/find*. Rogers (2019) states that by using this method, an MWU resource will end up with less redundancy and more accurate frequency counts. This is an improvement over studies that used an n-gramming approach, such as Simpson-Vlach and Ellis (2010), which resulted in redundant entries all with separate frequency counts in their resulting resource, such as:

there are a number of
there are a number
are a number of

Rogers (2017a) took the concgramming method further by creating a novel method to identify exemplary MWUs from lemmatised concgram data: First, concordance data from a corpus was collected for high-frequency co-occurring lemma with frequency and mutual information cutoffs. For example, *come* and *term* frequently co-occurred in Rogers' (2017a) initial data (see Table 1).

To identify the exemplary MWU of the data, any data that contained the core unit (*come to terms*) and occurred 50% or more of the core unit's frequency was considered the exemplary unit. The same method was applied if that unit could also be extended further. As seen in Table 1 above, *come to terms with* not only containing the core unit, but when the core unit occurs, it is also followed by *with* more than 50% of the time (229 times out of 243 occurrences). Data also indicates that this new exemplar occurs within *to come to terms with* more than 50% of the time as well (129 times out of 229 occurrences). However, the following extension (*to come to terms with the*) occurs less than 50% of the time as that current exemplar (44 times out of 129 occurrences) and thus, the extending stops there and *to come to terms with* is identified as the exemplary MWU of the lemmatised concgram *come/term*.

The significance of collocations

The significant role of collocations studies can be dated back to Rubin (1975) and Rubin and Thomp-

Table 1 MWUs identified from 500 example sentences in which the lemma pair “come” and “term” both occur (in data from Rogers. 2017a, p. 29).

MWU	Occurrences in 500 sentences
come to terms	243
come to terms with	229
to come to terms	133
to come to terms with	129
coming to terms	96
coming to terms with the	86
to come to terms with the	44
come to terms with [pre-nominal possessive pronoun]	28
coming to terms with the	26

son (1983), who defined some characteristics of Good Language Learners (GLLs). Rubin and Thompson stated that GLLs need to learn whole chunks of language in codified processes to help them perform beyond their capabilities. Also, the value of collocational knowledge is revealed by studies that found that a significant percentage of spoken discourse and written discourse consists of prefabricated chunks (Altenberg, 1998; Cowie, 1991, 1992; Erman & Warren, 2000; Foster, 2001; Hill, 2000; Jackendoff, 1997; Moon, 1998; Pawley & Syder, 1983; Ramisch et al., 2013). For instance, the investigation of the *Oxford Hector Pilot Corpus and Birmingham Collection of English Text* by Moon (1998) and one American television show by Jackendoff (1997) demonstrated that the two sources consisted of a high ratio of MWUs. Also, Pawley and Syder (1983) and Altenberg (1998) agreed that a high percentage of MWUs in spoken language existed.

As more researchers have recognised the significant role of MWUs, several empirical studies have been conducted to investigate them. These findings illustrated that having collocational knowledge results in faster reading, smoother output, quicker thinking, and the enhancement of communicative competence (Bahns & Eldaw, 1993; Conklin & Schmitt, 2008; Dechert, 1983; Hill, 2000; Howarth, 1998; Kuiper, 1996; Lewis, 1993, 2000; Nattinger, 1980; Underwood et al., 2004; Williams, 2002; Wray, 2000). Specifically, Lewis' (1993) lexical approach focuses on collocations and suggests that “increasing competence and communication power are achieved by extending the students’ repertoire of lexical phrases and collocational power” (p. 48). In addition, many collocations are associated with pragmatic functions such as politeness markers such as *I wonder if you mind* and discourse markers such as *let me see now* (Wray, 2000, p. 476). Therefore, obtaining such knowledge results in communicative competence. In brief, the knowledge of MWUs for L2 learners plays a central role in achieving fluency and efficient language processing. Since a significant component of linguistic proficiency is collocational knowledge, exploring the areas of difficulty in learning them could provide the first step to developing materials to help students better master them.

The effect of congruency on learning MWUs

A great deal of empirical studies has been done to investigate the areas of difficulty in learning MWUs (e.g., Laufer & Waldman, 2011; Nesselhauf, 2003; Wang & Shaw, 2008), the learning process of single words and MWUs (Alali & Schmitt, 2012; Laufer & Girsai, 2008), the relationship between collocation knowledge and language proficiency (Lee, 2015; Shokouhi & Mirsalari, 2010), and how L2 collocation processing occurs (Lee, 2016; Yamashita & Jiang, 2010). Also, there has been growing interest in identifying factors that affect the processing of L2 collocation, especially L1-L2 congruency (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2017) and semantic transpar-

ency (Gyllstad & Wolter, 2016; Yamashita, 2018).

Among these findings, several studies indicate that L2 learners make errors due to L1 influence (e.g., Bahns & Eldaw, 1993; Bahns, 1993; Davoudi & Behshad, 2015; Fan, 2009; Gyllstad, 2005; Martelli, 2006; Nakata, 2006; Phoocharoensil, 2011; Rogers & Florescu, 2016; Sen Bartan, 2019). More specifically, Nakata (2006, p.156) gives a few examples, such as how the terms 連絡をとる [renraku wo toru] and 犠牲を払う [gisei wo harau] can be mistakenly directly translated into English as *take contact* and *pay sacrifice*, respectively, while the correct English expressions are *make contact* and *make a sacrifice*. Rogers and Florescu (2016) also provided similar examples between Japanese and English, such as how the *setting sun* is the *sinking evening sun* in Japanese (沈む夕日 [shizumu yuuhi]).

In a similar vein, Davoudi and Behshad (2015) found that a lack of collocational knowledge was the reason for 75% of errors in essay writing by Iranian learners majoring in English language and English translation. The results of their study indicated that errors were rooted in L1 transfer (47.4%) and the use of synonyms (35.1%). For example, they found that the errors *bring some reasons* (state some reasons) and *learn knowledge* (gain knowledge) was rooted in L1 transfer. In a recent study, Sen Bartan (2019) showed that 47% of Turkish learners' verb + noun collocation errors were due to L1 influence. Thus, word-to-word translation or substituting one word with a synonym could result in unacceptable structures.

Furthermore, when reviewing the study by Boers (2020, p.146), the author claims that high-frequency transparent MWUs attract little attention (e.g., *do homework*). Contrastingly, learners may be puzzled by non-transparent collocations such as idioms (e.g., *over the hill*). Thus, when the meaning of collocation is transparent, learners may not notice the differences between the L1 and L2 collocation, and errors could occur in the production of MWUs. For example, if *do* in *do your homework* attracts little attention, the error *do a mistake* could occur due to L1 interference with Persian learners. Such examples of L1 interference support Shin and Chon's (2019, p. 610) view that L2 learners face “problems with all sort of MWUs of different degrees of compositionality (transparency).”

Collocations can thus be categorised into varying levels of congruency to help identify items with a higher learning burden. The results of studies by Yamashita and Jian (2010) for Japanese learners, Wolter and Gyllstad (2011) for Swedish learners, and Peters (2016) for Dutch learners all came to the same conclusion: These studies suggest that incongruent collocations are more challenging to learn than congruent collocations. In a follow-up study, Wolter and Gyllstad (2013) noted that the frequency of collocations in L2 is another factor that affects learning. In another study, Wolter and Yamashita (2015) confirmed the assumption that congruent collocations are learned quicker than non-congruent collocations because of positive L1 transfer.

Such findings have led to the evaluation of the congruency effect in the processing of L2 collocations becoming a more prominent topic in recent collocational research in various contexts to identify the relationship between the impact of L1 congruency with other factors (Du et al., 2021; Fang & Zhang, 2021; Özdem-Ertürk, 2021; Sonbul & El-Dakhs, 2020). For instance, Özdem-Ertürk (2021) investigated interlexical factors (congruency with L1) and intralexical factors (collocational frequency, node word frequency, mutual information score, and type of collocation) on receptive and productive Turkish collocational knowledge. The researcher found that, among these five factors, both receptive and productive collocational knowledge were affected by congruency in the L1. In another study, Sonbul and El-Dakhs (2020) focused on the interaction between the effect of congruency and the level of proficiency of Arab learners of English. They found that these two factors affected a timed and untimed collocation recognition test. Thus, the congruency effect was reduced by increasing the level of proficiency.

L2 learners need to get more input or more frequent encounters with incongruent rather than congruent collocations to reduce errors. Also, L1-L2 congruency could be a fundamental criterion in creating collocation resources to help the learner focus on incongruent items and avoid errors. However,

more research still needs to be done as such investigations are complex and time-consuming. For instance, Shin (2006) stated that L1-L2 congruency is a fundamental factor to consider, but he only checked 500 of the approximately 4,500 items in his study. Rogers (2017a) was more thorough in his list of approximately 11,000 high-frequency MWUs with L1-L2 contrastive analysis revealing that around half of the English items examined were incongruent with Japanese (L1). He concluded that the high ratio of incongruent collocations made it clear that L1-L2 congruency was essential when choosing English MWUs for Japanese learners to focus on. Therefore, congruency could be considered a fundamental criterion to help select useful MWUs for L2 learners.

Contrastive analysis

The role of the L1 in the learning process originated in the 1940s via behaviourism, which emphasised learning as a habit formation process. From this perspective, L2 learning occurs by replacing L1 habits with new habits in the L2 context, and L1 interference was recognised as a significant problem in the learning process. One of the well-known theories that recognise the effect of L1 on L2 is contrastive analysis, which is rooted in behaviourism. The contrastive analysis focuses on the similarities and differences between the learner's L1 and L2 to predict what challenges a learner of a specific L2 will encounter. This term was developed by Lado (1957), who claimed that the degree of difficulty in learning a second language is connected to the degree of difference between the two languages.

However, as behaviourism began to be rejected and generative linguistics became more prominent, contrastive analysis dramatically declined because not all errors are related to L1 interference or language transfer. On the other hand, other areas of focus emerged, such as error analysis, cross-linguistic influence, and the role of L1 in the cognitive approach, which can be considered evidence to confirm the effect of L1 on learning an L2. In other words, different perspectives indicate L1 interference in learning L2 without needing to accept behaviourism (Thao, 2020).

Despite criticism of contrastive analysis, some researchers believe that it is a practical approach for comparing languages, translating, creating bilingual materials, explaining learners' errors, identification of difficulties, and exploring the fundamental aspects of the language learning process (Ellis & Barkhuizen, 2005; Keshavarz, 2011; Laufer & Girsai, 2008; Nesselhauf, 2005; Tajareh, 2015). Therefore, contrastive analysis can be desirable (Nesselhauf, 2005) and is considered by some researchers to be essential in teaching (Laufer & Girsai, 2008). Ellis and Barkhuizen (2005) stated that contrastive analysis serves two purposes: first, it explains the learners' errors; second, it acts as a data source to define the systemic areas that the target language teachers should consider. The proponents of contrastive analysis seem to believe it is "the main criterion for the preparation of instructional materials" (Keshavarz, 2011, p. 9).

The purpose of different approaches is to address learners' needs, and each approach has its advantages and disadvantages. Thus, the choice of method depends on the purpose of learning (Barghamadi, 2020). Since translation and creating bilingual materials can be aided by contrastive analysis, and there has been considerable agreement among researchers about the effect of L1 on learning MWUs, the current study utilised contrastive analysis. Our focus on Persian learners is similar to other approaches that have been taken thus far. For example, Sen Bartan (2019) suggested that it would be beneficial to conduct a contrastive analysis and prepare bilingual lists of lexical collocation errors specific to Turkish learners of English to explore the frequency of these errors. In the same vein, Berti and Pinnavalia (2012, p. 216) argued that the advantage of a bilingual dictionary for Italian learners is that it helps them search for a particular collocation. Also, in a case study, Ziafar (2015) concluded that Persian learners are greatly challenged when trying to figure out what formulaic expressions mean because there is no organised resource for learning them. Since there is a need for a comprehensive resource focusing on English-Persian congruency, this study aims to fill this gap in the research.

Methodology

In this research, Rogers' list (2017a) of high-frequency English MWUs is utilised to identify items with low L1-L2 congruency with Persian to create a resource to help Persian-speaking learners teach

items with low L1-L2 congruency. To develop the list of 11,212 MWUs, Rogers (2017a) used Davies' (2010) *Word List Plus Collocates* as a starting point for collocations to search for and analyse corpus data containing them in the *Corpus of Contemporary American English* (COCA) (Davies, 2008). To delimit items to only the most common, he set a cut-off of one occurrence per million tokens, utilised mutual information data to delimit collocations identified, and only included items with balance dispersion and chronological data. Then, the custom one-off software *AntWordPairs* (Anthony, 2013) was developed for his study to analyse 500 examples of sentences that contain each lemmatised collocation to identify the most frequent MWU the lemma in.

Rogers' (2017a) list was chosen as a starting point in this study because it is the only large-scale lemmatised MWU list created in a way that considers essential factors, such as avoiding duplicate entries by using the lemmatised conprogramming approach. This list was also utilised because this resource was adopted by an entire Japanese foreign language faculty department as an official part of required study for all 1st and 2nd-year students. In addition to being used in this capacity for 6 years, data indicated that students who studied the contents for the most hours had the highest TOEIC score gains (Rogers, 2017b). Moreover, this study also found that students who received the highest quiz grades on the resource correlated highly with TOEIC grades in that such students achieved the highest TOEIC scores in the faculty.

A novel approach was used to create this research; thus, there are still limitations regarding confirming its validity. However, all available data indicate that it has value for second language learners and that its approach has advantages over previous studies. Therefore, it was chosen as a starting point for the current research to validate its usefulness further. Thus, this current study will use his list of 11,212 MWUs as a starting point for L1-L2 contrastive analysis with Persian. This research used a 12-point system for L1-L2 (Persian/English) congruency ratings following Rogers' (2017a) own approach for comparing the list with Japanese. In this system, 12 points were given for total congruency when each component of MWUs was equal to the Persian translation of the collocation. For example, when MWUs consisted of 2 words and had the same meaning in both Persian and English, each word received 6 points, thus receiving 12 points in total. If one of the words did not match, the score was 6. For example, *take a photo* is literally the same as the Persian meaning (عکس گرفتن). Thus, this MWU received 12 points.

On the other hand, *make a mistake* (عکس گرفتن) is not congruent to Persian and thus received a rating of 6 points because the verb *make* is not congruent to the Persian meaning. The rating system consisted of 0, 3, 4, 6, 8, 9, and 12 points, and all the points were rounded to these numbers when the number of MWUs was more than 4 words long to create a balanced system to analyse data later easily. Since the base list was an American corpus, the translation process relied upon the online *Longman Dictionary* and *Merriam-Webster American English Dictionary* as references to double-check the meaning of some MWUs in the context that has idiomatic meaning (e.g., *a leg up on, set the record straight*).

The following protocol was used for rating the MWUs:

- If a word did not match but was within the same word family, it received half its allotted points.
- The articles *a/an/the* do not exist in Persian and thus were ignored in the rating process.
- Any differences in word order were ignored.
- Loan words used in Persian, such as *a coffee shop*, were considered congruent.

L1-L2 congruency ratings were given to each MWUs in the list by the first author with native-like abilities in both languages. A second rater was asked to rate 10 per cent of target items to ensure the protocol was reliable.

Results and Discussion

There is consensus among researchers that L1-L2 congruency affects the learning burden of MWUs (e.g., Nesselhauf, 2005). Accordingly, this study set out to give L1-L2 ratings to 11,212 English MWUs and analyse their Persian translations to identify items with low L1-L2 congruency. The full

Table 2 *L1-L2 congruency ratings of high-frequency English MWUs with Persian translations*

Congruency Rating	Count	Percentage
0	498	4.4
3	260	2.3
4	1,026	9.2
6	2,886	25.7
8	1,654	14.8
9	1,070	9.5
12	3,818	34.1
Total	11,212	100.0

results can be viewed via this link: <https://jamesmartinrogers.wixsite.com/mwusforpersians>. Table 2 is a summary of these results. Based on the 12-point system, MWUs with ratings under 12 points were categorised as incongruent to some extent. Therefore, the data shows that a majority (65.9%) of the items were considered either somewhat incongruent or totally incongruent with their Persian translations. Table 2 shows that 3,818 (34.1%) of MWUs are rated in 12 points; thus, these items in Rogers' list are congruent with Persian translation.

According to the literature review, incongruent collocations require more study time due to the higher learning burden. A list of items with a higher learning burden can be arranged according to L1-L2 congruency, with more challenging items requiring additional study time or teaching. In this study, 7,394 items were found to be incongruent, to some extent, with Persian. Learners could achieve fluency more efficiently if the volume of learning items were reduced, so a cut-off of 6 out of 12 points was used. As a result, the 11,212 items were reduced to 4,670 items (41.6%), half of which significantly differ in their translation. Such a list would be particularly beneficial for learners with limited time to study but who wish to concentrate only on items likely to cause errors. Table 3 provides a sample of every 100th MWUs identified as incongruent items with their Persian translations in this study .

The interrater reliability results showed nearly 96% agreement between the two raters. Since researchers have deemed interrater reliability to be from 75% to 90% (Hartmann, 1977; Stemler, 2004), these results point to the protocol for assigning L1-L2 congruency ratings being highly reliable.

It is widely recognised that collocations are essential when learning languages. The evidence demonstrated that learning an L2 involves noticing lexical chunks that repeat in learners' input (Lewis, 1993). Collocational knowledge constitutes a substantial part of linguistic proficiency, so exploring areas of difficulty could provide a guide for developing materials. Since several studies show that L1 interference is the main factor in learning these items, considering L1-L2 concurrency in developing collocations resources could be a reasonable approach.

The main goal of this study was to determine whether or not L1-L2 congruency could be a critical criterion for identifying high-frequency MWUs that need extra study time. An L1-L2 contrastive analysis was conducted on 11,212 English MWUs identified by Rogers (2017a). The results indicated that a majority (65.9%) of the MWUs in Rogers' list were not congruent to some extent with their Persian counterparts. Thus, this criterion was shown to be an essential factor to consider when selecting high-frequency MWUs for direct study. As discussed in the literature review, Shin (2006) and Rogers (2017a) conducted a contrastive analysis to examine L1-L2 congruency. Since this task is very time-consuming, Shin only checked 500 of the approximately 4,500 items in his study. Rogers (2017a) was significantly more comprehensive. Therefore, only one major work in the literature serves as a direct point of comparison for the current study. There was a difference in the results of his research and this current study to note: Rogers found that 30.5% of the items received a rating of 0-6 for English-Japanese, while this current study found 41.6% of the items rated 0-6 for English-

Table 3 Sample of MWUs with low L1-L2 rating

Frequency	Collocate	Pivot Word	MWU
11986	go	off	Went off
7933	off	cut	cut off from
5851	give	away	to give away
4630	get	little	to get a little
3941	come	around	come around
3450	away	throw	throw away
3034	out	pass	passed out
2692	trade	agreement	trade agreement
2441	up	screw	screwed up
2260	research	project	research project
2095	material	raw	the raw materials
1945	chairman	committee	committee chairman
1813	leave	second	seconds left
1722	great	depression	the great depression
1632	down	crack	crack down on
1554	almost	entirely	almost entirely
1477	make	run	to make a run for
1403	rest	country	the rest of the country
1343	go	forth	go back and forth
1283	university	public	public university
1235	information	personal	personal information
1198	fight	off	to fight off
1159	television	program	television program
1115	charge	file	file charges
1082	drag	out	dragged out
1050	table	side	side of the table
1020	perspective	put	put it in perspective
994	race	ethnicity	race and ethnicity
969	policy	security	national security policy
947	political	correctness	political correctness
915	hold	tight	hold tight to
889	area	remote	in remote areas
860	financial	assistance	financial assistance
833	birth	certificate	birth certificate
811	way	positive	in a positive way
786	animal	species	plant and animal species
763	bear	child	bear children
739	home	abroad	at home and abroad
716	fail	meet	failed to meet
715	interest	payment	interest payments
694	make	visible	made visible
667	667	clue	give you a clue
640	ahead	week	in the weeks ahead
618	concern	growing	a growing concern
600	camera	crew	a camera crew
578	issue	issue	issues that arise
555	process	healing	the healing process
515	medicine	medicine	alternative medicine

Persian. Thus, there were more incongruences between Persian and English in comparison with Japanese.

The overall results showed that L1-L2 congruency is a significant criterion to consider when identifying MWUs that deserve extra study time because, as Gyllstad & Wolter (2016) note, the most significant barrier in L2 collocational processing is the lack of congruency between the L1 and the L2. Other researchers agree that when looking at L1-L2 investigations, learning congruent collocations is easier and quicker than learning non-congruent (Bahns, 1993; Gyllstad & Wolter, 2016; Nesselhauf, 2005; Wolter & Gyllstad, 2011, 2013; Yamashita & Jiang, 2010) and learners make more errors with incongruent collocations (Davoudi & Behshad, 2015; Yamashita & Jiang, 2010).

The lack of congruence between L1 and L2 appears to have the most significant impact on learning. If the vast majority of collocations are incongruent to some degree, then emphasis on L1-L2 congruency would be reasonable. Based on this study, we recommend that L1-L2 congruency be utilised as a criterion in English language learning materials development for Persian-speaking learners. Due to the higher learning burden of incongruent items, it is beneficial to list items needing additional study time. Moreover, a list like this could also be helpful for learners with limited study time to enable them to focus more on items they have a higher chance of making an error with. Thus, this study's results could be a valuable resource to help Persian learners develop their collocational fluency. Further, increasing learners' exposure to MWUs and focusing their attention on particular MWUs to gain fluency leads to an emphasis on explicitly taught items (see Barghamadi et al., 2022). Therefore, teachers may find it helpful to follow some guidelines based on the findings of this study:

- MWUs should be explicitly taught.
- Introduce congruent and incongruent types of collocations to learners.
- Create custom-tailored learning resources for their L1.

This evidence could emphasise the need for custom-tailored materials for specific learner groups compared to generic second language learning materials designed for any learner. Consequently, creating bilingual materials can be aided by contrastive analysis. For example, this study's findings could be integrated into a curriculum as supplementary study materials via Leitner algorithm-based spaced repetition software, as Rogers (2017b) did for a foreign language department in a Japanese university. His study found a high correlation between knowledge of the MWUs and high TOEIC scores. He also found a high correlation between study time using the software and TOEIC score increases. Similar studies could be conducted with Persian-speaking learners to determine if such findings can be replicated. This study's results could also be used as reference material for textbook writers who intend to create materials that target Persian-speaking English learners. For example, the reading passages in an English reading textbook for Persian speakers could introduce and highlight the MWUs and include their Persian translations in the textbook's margin. As such a textbook series advances to higher levels, the author can utilise the frequency organisation of this current study's resource to introduce more challenging items.

Conclusion

This study investigated the level of L1-L2 congruency between the Persian translation equivalents of high-frequency MWUs. It was found that the majority of the MWUs examined were incongruent to some extent with their Persian translations. Because research indicates that incongruent items lead to a higher learning burden, the finding of this study demonstrates that L1-L2 congruency is a significant factor between English and Persian, and consideration for it has the potential to increase the efficacy of English language acquisition among this learner group.

While the substantial number of the words in this study were incongruent, the study's results should be interpreted cautiously due to limitations. First of all, the method utilised to identify the 11,212 MWUs used as a starting point in this study is new and novel; thus, there is no way to compare it to other studies to validate it. In addition, there is a subjective element to the rating process, albeit one that cannot be avoided. Moreover, because of the sheer size of the study, it was only possible to use one translator. Ideally, the translations would be checked by other translators, have multiple raters,

and interrater reliability analysis would be conducted for the entire list. However, interrater reliability was very high at 96% for the items rated by multiple individuals. Thus, future research requires validating the study's starting point of using the 11,212 MWUs and confirming the translations and ratings. This study is a first step toward creating a resource that can help Persian learners improve their English collocational fluency with better efficacy.

References

- Alali, F. A., & Schmitt, N. (2012). Teaching formulaic sequences: The same as or different from teaching single words? *TESOL Journal*, 3(2), 153–180. <https://doi.org/10.1002/tesj.13>
- Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101–122). Oxford University Press.
- Anthony, L. (2013). AntWordPairs [Computer software] Version 1.0.2 (Available on request). <https://www.laurenceanthony.net/software>
- Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal*, 47(1), 56–63.
- Bahns, J., & Eldaw, M. (1993). Should we teach ESL students collocations? *System*, 21(1), 101–114.
- Barghamadi, M. (2020). A critical study of the principles and approaches to needs analysis. *Studies in Educational Management*, 7, 1–16. <https://doi.org/10.32038/sem.2020.07.01>
- 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>
- Benson, M., Benson, E., & Ilson, R. (1986). *Lexicographic description of English*. John Benjamins Publishing Company.
- Berti, B. & Pinnavalia, B.B. (2012). Towards a corpus-driven bilingual Italian-English dictionary of collocations. In R. Facchinetti (Ed.), *English Dictionaries as Cultural Mines* (pp. 201–223). Cambridge Scholars Publishing.
- Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 143–157). Routledge. <https://doi.org/10.4324/9780429291586-10>
- Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4), 411–433.
- Cowie A. P. (1992). Multiword lexical units and communicative language teaching. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 1–12). Macmillan.
- Cowie, A. P. (1991). Multiword units in newspaper language. In S. Granger (Ed.), *Perspectives on the English lexicon: A tribute to Jacques Van Roey* (pp. 101–116). Cahiers de l'Institut de Linguistique de Louvain.
- Cowie, A. P. (Ed.). (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.
- Davies, M. (2008). *The corpus of contemporary American English: 425 million words, 1990–present*. Brigham Young University. <http://corpus.byu.edu/coca/>
- Davies, M. (2010). *Word list plus collocates*. <http://www.wordfrequency.info/purchase1.asp?i=c5a>
- 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/tpis.0512.29>
- Du, L., Elgort, I., & Siyanova-Chanturia, A. (2021). Cross-language influences in the processing of multiword expressions: From a first language to second and back. *Frontiers in Psychology*, 12, 666520. <https://doi.org/10.3389/fpsyg.2021.666520>
- Durrant, & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157–177. <https://doi.org/10.1515/iral.2009.007>
- Durrant. (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>
- Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.
- Ellis, R. & Barkhuizen, G. (2005). *Analysing learner language*. 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.
- Fan, M. (2009). An exploratory study of collocational use by ESL students – A task-based approach. *System* 37(1), 110–123.
- Fang, N., & Zhang, P. (2021). L1 congruency, word frequency, collocational frequency, L2 proficiency, and their combined effects on Chinese–English bilinguals' L2 collocational processing. *International Journal of Bilingualism*, 1–17. <https://doi.org/10.1177/13670069211024747>
- Fellbaum, C. (2015). The treatment of multi-word units in lexicography. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 411–424). Oxford University Press.
- 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.
- 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>
- Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal of Applied Behavior Analysis*, 10, 103–116.
- Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp.47–67). Language Teaching Publications.
- Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24–44.
- Hunston, S., & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing Company.
- Jackendoff, R. (1997). “Twistin” the night away. *Language*, 73, 534–559.
- Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Corpus linguistics and beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora* (pp. 133–140). Rodopi.
- Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Clarendon Press.
- Kuiper, K. (1996). *Smooth talkers*. Lawrence Erlbaum.
- Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.
- Laufer, & 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, & 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>
- Lee, S. (2015). Korean EFL learners' productive collocation knowledge and its relationship to their writing proficiency. *English Language and Literature Teaching*, 21(4), 59–84.
- Lee, S. (2016). L1 influence on the process of L2 collocation: An experimental study of Korean EFL learners. *Linguistic Research*, 33, 137–163.
- Lesniewska, J. & Witalisz, E. (2007). Cross-linguistic influences on L2 and L1 collocations. *EU-ROSLA Yearbook*, 7, 27–48.
- Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Language Teaching Publication.
- Lewis, M. (Ed.). (2000). *Teaching collocation: Further developments in the lexical approach*. Language Teaching Publications.
- Liu, E. T., & Shaw, P. M. (2001). Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics in Language Teaching*, 39(3), 171–194.
- Martelli, A. (2006, September). A corpus-based description of English lexical collocations used by Italian advanced learners. In *Atti del XII Congresso Internazionale di Lessicografia* (pp. 1005–1011).

- Masini, F. (2019). Multi-word expressions and morphology. In the *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780199384655.013.611>.
- Moon, R. (1998). *Fixed expressions and idioms in English. A corpus-based approach*. Oxford University Press.
- Nakata, T. (2006). English collocation learning through meaning-focused and form-focused activities: Interactions of activity types and L1-L2 congruence. In the *Proceedings of the 11th Conference of Pan-Pacific Association of Applied Linguistics* (pp. 154–168).
- Nation, I. S. P. (2001). *Learning vocabulary in another language* (1st ed.). Cambridge University Press.
- 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 Webb, S. (Ed) *The Routledge Handbook of Vocabulary Studies* (pp. 15–29). Routledge.
- Nattinger, J. R. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly*, 337–344.
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223–242.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Publishing Company.
- Özdem-Ertürk, Z. (2021). *Factors affecting and productive knowledge of collocations of tertiary level learners of English in Turkey* (Unpublished doctoral dissertation). University of Hacettepe University, Turkey.
- Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.
- Peters. (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>
- Ramisch, C., Villavicencio, A., & Kordoni, V. (2013). Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing*, 10(2), 1–10. <https://doi.org/10.1145/2483691.2483692>
- Rogers, J. (2017a). *What are the collocational exemplars of high-frequency English vocabulary? On identifying MWUs most representative of high-frequency lemmatized concgrams* (Unpublished doctoral dissertation). University of Southern Queensland, Australia.
- Rogers, J. (2017b). On the implementation of a smartphone-based English app study program. *The Proceedings for the First International Forum* (pp. 71–76). Nagoya: Faculty of Foreign Studies, Meijo University.
- Rogers, J. (2019). Conceptualizing collocations and multi-word units as the same through the concgramming approach (Special issue: A dedication of appreciation and recognition of service by Dr. Yasumi Murata, Professor in the Faculty of Foreign Studies, Meijo University). *名城大学外国語学部紀要*, 3, 147–151.
- Rogers, J., & Florescu, C. (2016). On L1-L2 congruency as a criterion for the selection of multi-word units worthy of instruction. *名城大学人文紀要*, 52(2), 31–37.
- Rubin, J., & Thompson, L (1982). *How to be a more successful language learner*. Heinle & Heinle.
- Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 1–22). John Benjamins Publishing Company.
- Şen, B. Ö. (2019). Lexical collocation errors in literary translation. *Dil Dergisi*, 1(170), 73– 88.
- Shin, D. (2006). *A collocation inventory for beginners*. (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.
- 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>
- Shokouhi, H., & Mirsalari, G.-A. (2010). Collocational knowledge versus general linguistic knowledge among Iranian EFL learners. *TESL-EJ*, 13(4), 24.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512. <https://doi.org/10.1093/applin/amp058>
- Snellings, P., Van Gelderen, A., & De Glopper, K. (2002). Lexical retrieval: An aspect of fluent second-language production that can be enhanced. *Language Learning*, 52(4), 723–754. <https://doi.org/10.1111/1467-9922.00202>

- 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(4). <https://scholarworks.umass.edu/pare/vol9/iss1/4>
- Tajareh, M. J. (2015). An overview of contrastive analysis hypothesis. *Cumhuriyet Üniversitesi Fen Edebiyat Fakültesi Fen Bilimleri Dergisi*, 36(3), 1106–1113.
- Thao, M. T. P. (2020). Role of L1 in L2 acquisition according to contrastive analysis hypothesis and error analysis. *Journal of Education and Practice*, 11(2), 102–107. <https://doi.org/10.7176/JEP/11-2-11>
- Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp. 153–172). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.9.09und>
- Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME journal*, 32, 201–232.
- Williams, G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1), 43–64.
- Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32(4), 430–449. <https://doi.org/10.1093/applin/amr011>
- 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>
- Wolter, B., & Yamashita, T. (2017). Word frequency, collocational frequency, IA congruency, and proficiency in L2 collocational processing: What account for L2 performance? *Studies in Second Language Acquisition*, 40(2), 395–416. <https://doi.org/10.1017/S0272263117000237>
- Woolard, G. (2000). Collocation – Encouraging learner independence. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 28–46). Language Teaching Publications.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics*, 21(4), 463–489.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.
- 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>.
- Ziafar, M. (2015). Good language learners and their use of lexical items and L1 in second language acquisition. *International Scientific Research Journal*, 1(5), 1–9. <https://doi.org/10.18483/IRJSci.1>