How Much Collocation Knowledge Do L2 Learners Have?

The Effects of Frequency and Amount of Exposure

Abstract

Many scholars believe that collocations are difficult to learn and use by L2 learners. However, some research suggests that learners often know more collocations than commonly thought. This study tested 108 Spanish learners of English to measure their productive knowledge of 50 collocations, which varied according to corpus frequency, *t*-score, and *MI* score. The participants produced a mean score of 56.6% correct, suggesting that our learners knew a substantial number of collocations. Knowledge of the collocations correlated moderately with corpus frequency (.45), but also with everyday engagement with English outside the classroom, in activities like reading, watching movies/TV, and social networking (composite correlation = .56). Everyday engagement also had a stronger relationship with collocation knowledge than years of English study (.45).

Keywords: collocations, productive knowledge, frequency, exposure, acquisition

Introduction

It is now well established that formulaic language provides processing advantages and is essential for using language fluently and idiomatically, both for native and non-native speakers (for overviews, see Schmitt, 2004; Wray, 2002, 2008, and *Annual Review of Applied Linguistics 32*, 2012). Further, it has been suggested that if second language learners aim to achieve nativelike mastery, they need to know formulaic language and use it accurately and appropriately (Ellis *et al*. 2008; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983). However, despite the importance and value of learning such formulae, research has demonstrated that it is a difficult aspect for even advanced L2 learners (Granger, 1998; Howarth, 1998; Nesselhauf, 2003, 2005), and that their knowledge of formulaic sequences lags behind their general language and vocabulary knowledge (Bahns and Eldaw, 1993; Irujo, 1993). Hence, even though formulaic language has been found to be important for the processing, production and acquisition of natural language, there are still many questions about how formulaic language is acquired by L2 learners.

It has been proposed that extensive exposure is a key factor necessary for this acquisition (Nattinger and DeCarrico, 1992; Durrant and Schmitt, 2010; Martinez and Schmitt, 2012). Presumably this extensive exposure will largely be driven by the frequency of each formulaic sequence in naturally-occurring language, with more frequent items generally being better learned. It is widely recognized that individual

words respond to the effects of frequency, so that learners generally acquire higher-frequency words before lower-frequency ones (Nation and Waring 1997; Leech *et al*., 2001; Nation, 2001; Ellis, 2002). However, can this finding for single words also be applied to formulaic language, specifically collocations? Furthermore, although frequency counts are currently derived from very large corpora such as the COCA (Corpus of Contemporary American English – 450 million words (Davies, 2008)) or the BNC (British National Corpus - 100m), no corpus can replicate the exposure any individual person has, especially L2 learners. So the question remains whether the approximate exposure information (i.e. frequency) available from corpora can indicate learner knowledge of formulaic sequences. Perhaps an equally useful predictor is the degree to which learners engage with and use the L2 (e.g. when studying an L2, or using an L2 in social networking or watching films and television)? This study will explore how both collocation frequency and measures of language usage relate to knowledge of collocations by Spanish learners of English.

A Focus on Collocations

Approaches to Definition and Identification

Formulaic language consists of a number of different categories, each with their own characteristics, behaviors, and problems, such as idioms, lexical bundles, phrasal verbs, and phrasal expressions. But perhaps the most studied category in applied linguistics is collocation.[1] Collocation refers to the idea of lexical patterning, and Schmitt's (2000: 76) definition is typical: "the tendency of two or more words to co-

occur in discourse". Broad definitions of collocation such as this have been operationalized through two main approaches. The first is phraseological, where collocations are seen as word combinations with various degrees of fixedness (e.g. Cowie, 1998). The second more common approach is statistical, where various formulas are used to search corpora and identify the words which pattern together (e.g. Sinclair, 2004).

However, different statistics can identify quite different sets of collocations, although there will always be some overlap. Three commonly used statistical measures are frequency, *t*-score, and *MI*. Frequency of occurrence in a corpus identifies collocations which are common and are often meaningful across a wide range of contexts (e.g. *last night* - 17,214 instances in the COCA, *long ago* – 8,455, *hard work* – 4,763). *T*-score weights frequency highly in its calculations, and so also identifies collocations which are frequent in use. *MI* identifies collocations which are typically not very frequent, but which have a particularly strong bonding when they do occur (e.g. *commit suicide*, *resist temptation*). Unfortunately, it is not clear which of these (or other) measures is the best to use in research, and to date, the selection of one or another seems to be somewhat arbitrary. Knowing how these methods relate to learner knowledge might be helpful for researchers when selecting a method.

L2 Knowledge and Use of Collocations

Formulaic language in general is widely employed by native-speakers, and provides many processing and communicative advantages (e.g. *Annual Review of Applied*

Linguistics 32, 2012). The same is true for collocations more specifically, and it seems clear that knowledge of collocations can greatly benefit second language learners in their attempts to achieve high proficiency in an L2. Indeed, lack of collocation knowledge has been shown to be problematic. Receptively, this lack can lead to miscomprehension (Martinez and Murphy, 2011).[2] Productively, the lack of use of collocations, as well as the over-, under- or mis-use of them, lead to L2 speakers being judged as odd, unnatural or non-nativelike (e.g. Granger, Paquot, & Rayson, 2006; Barfield & Gyllstad, 2009), while use of formulaic sequences is related to higher ratings of learner proficiency (e.g. Boers, *et* al., 2006). Therefore, it is widely accepted that if L2 learners want to use language accurately and fluently, they need to know and use collocations.

But how many collocations do L2 learners use, and how well? In recent decades, research has begun to address these questions. Unsurprisingly, it has been shown that knowledge of collocations by L2 learners is lower than that of native speakers, with L2 learners often misusing these combinations and making many mistakes (e.g. Granger, 1998; Bahns and Eldaw, 1993; Howarth, 1998; Laufer & Waldman, 2011). Some researchers believe the knowledge of collocations by L2 learners lags behind their general language and vocabulary knowledge (Bahns and Eldaw, 1993; Wray, 2002). Others, like Farghal and Obeidat (1995) believe that, in general, L2 learners do not know collocations because they were not made aware of them, so they have to make use of strategies, such as the use of synonyms, paraphrasing or avoidance. However, just asking learners to pay attention to collocations does not seem to be that effective, and focusing learners' attention on target items seems to require some explicit approach, such as using typographical marking, e.g. bold and underlined font

(Peters, 2012). Regardless of the underlying cause, L2 learners seem to lack the range of collocations available to natives.

However, the evidence also shows that L2 learners can produce numerous correct collocations. For example, Siyanova and Schmitt, (2008) extracted 810 adjective–noun collocations from the ICLE sub-corpus written by Russian learners of English (Granger, Meunier, Paquot, n.d.), and found that around 45% were appropriate, based on frequency and Mutual Information (*MI*) scores. When the L2 results were compared to those from native speakers, very little difference was found based on these criteria. Nesselhauf (2005) also found that German learners of English made extensive use of collocations in her corpus research. However, she also reported that they often used them in an inappropriate manner, which suggests that what is problematic for L2 learners regarding collocation use is not so much knowing which words co-occur, but rather when and how to employ those combinations appropriately. This is congruent with research which shows that L2 learners tend to adhere to and overuse some collocations they feel more confident with (what Hasselgren (1994:237) called “teddy bears”), and underuse native-like collocations, making use of atypical, idiosyncratic ones (Granger, 1998; Howarth, 1998; Lorenz, 1999; De Cock, 2000). Laufer and Waldman (2011) found that, independently from their proficiency level, learners produced non-standard, idiosyncratic collocations when compared to natives. Similarly, Millar (2011) showed that L2 learners made more use of non-conventionalized collocations, and that, when this happened, natives required more time to process them.

This contradiction between learners having problems with collocations, while at the same time demonstrating the ability to use many appropriately, can partially be explained by the fact that collocations are not all the same. There is some evidence

that learners know the kind of collocations identified by frequency/*t*-score better than those identified by *MI*. Durrant and Schmitt (2009) analyzed a corpus composed of written academic output from Turkish and Bulgarian university EFL students and a mixed group of international university students studying in the UK. They found that these students tended to use frequent premodifier-noun collocations identified by *t*-score (*good example, long way*) at a rate similar to native students. However, the learners produced many fewer *MI* collocations (low-frequency but tightly-bound, e.g. *densely populated* and *preconceived notions*). Because of their strong ties, and relative infrequency, *MI* collocations are likely to be especially salient for natives, and so their absence in nonnative output is particularly noticeable. The authors conclude that the lack of these *MI* collocations is one key feature which distinguishes native from nonnative production.

In sum, research on the collocations used by L2 learners found that, even though learners *do* use them, they differ in their use compared to natives, both in quantity (using fewer of them) and quality (diversity, accuracy and appropriacy). However, this research has focused on learner output (quite often written compositions), which cannot indicate the whole range of collocations a learner might know. Alternatively, studies have used a small set of collocations which were chosen for their particular characteristics (e.g. opaqueness or syntactic construction (e.g. Adjective+Noun)), and may not be representative of the larger population. Neither approach really had the purpose (or the methodology) to obtain a measurement of overall collocation knowledge. Thus, we are still left with the question of how much collocation knowledge L2 learners have. That is, do learners really have a quite limited knowledge of collocations as many scholars suggest, or do they actually know a

wider range of collocations which has not been captured by previous research methodologies?

The Acquisition of Collocations

The Role of Frequency in the Acquisition of Collocations

Usage-based theories of language maintain that frequency of co-occurrence of linguistic items in the input is the main determining factor of the acquisition of these items. Therefore, knowledge of a language is related to the language exposure and the frequency of use of specific constructions (Ellis, 2002). Psycholinguistic research has demonstrated that both native and nonnative speakers are sensitive to the frequency of a wide range of linguistics forms, from phonemes to formulaic language (Bybee and Hopper, 2001; Ellis, 2002). In the case of vocabulary, it is thought that learners tend to know high-frequency lexical items better because they encounter them more often. In fact, frequency of occurrence has been widely recognized as one of the best predictors of usefulness and acquisition of individual words (Leech *et al.* 2001; Nation, 2001; Schmitt, 2010). According to Nation and Waring (1997), there is no reason to believe that formulaic sequences like collocations would behave differently in this respect.

Indeed, L2 learners seem to be able to acquire and use the collocations which appear frequently, but do not seem to pick up as many non-frequent collocations, whose individual component words may also be infrequent in themselves. This is highly suggestive of the role of frequency in the acquisition process. At the very initial stage of learning, Durrant and Schmitt (2010) found that even one exposure to a

word combination led to a small, but significant, facilitation of collocation completion in a priming experiment. However, they found that two repetitions of word combinations led to a large facilitation effect. In a study of incidental acquisition from a graded reader, Taiwanese university students generally learned more collocations as the frequency of exposure increased up to 15 (Webb, Newton, and Chang, 2013). Webb (2007) exposed learners to nonwords in reading texts once, three times, seven times, and 10 times. He found that 10 exposures led to significant gains in both receptive and productive knowledge of collocation. Likewise, Peters (2014) found that increased repetition in an explicit learning task improved learning of the target collocations. These results are indirectly supported by a study into 3-, 4-, and 5-word academic formulas, where Ellis, *et al* (2008) found that for natives, it is predominately the *MI* of a formulaic sequence which determines processability, while for nonnatives, it is predominately the frequency.

So it seems that frequency of exposure does affect the acquisition of collocations to some extent. This is fine in tightly-controlled experiments, but in most learning contexts it is impossible to either know or control for the number of exposures any learner receives. This leaves corpora as the main indicator of frequency. But does corpus frequency really indicate the likelihood of collocation acquisition? There are some reasons to think not. Assuming that the frequency of collocations in a corpus like the COCA represents the input received by learners is unrealistic. The input that those participants have been most exposed to is the language of instruction in a classroom. That language is largely different from that which can be found in a natural native environment, making expressions that are not that frequent in natural language much more frequent in the classroom, and vice versa. Moreover, formulaic

language in general, and collocations in particular, have been claimed to present a challenge for even very advanced L2 speakers. As the majority of ELT teachers are (hopefully advanced) L2 speakers themselves, their students' exposure to collocations can be limited (Meunier, 2012). Even when they use collocations, they may not use them like natives do, overusing some that are well-known to them even if they are not that frequent in native language.

So the question remains of how well frequency, as indicated by large native corpora, relates to the collocation knowledge of L2 learners. We found only one study where corpus frequency and knowledge of formulaic language were directly compared. Schmitt and Redwood (2011) investigated the learners' productive and receptive knowledge of phrasal verbs and the effect of their frequencies on this knowledge. They found that, for the productive test, frequency only explained 20% of the variance (r^2) in the scores when correlated with the frequency scores taken from the BNC, and 18% when compared to the frequencies of those phrasal verbs in the COCA. For receptive knowledge, the co-variance was even lower at 9% (BNC) and 13% (COCA). Overall, there was a general trend of higher frequency leading to a greater chance of learning phrasal verbs to productive degree of mastery. The relationship was not strongly linear, but higher frequency phrasal verbs were clearly learned by a greater number of participants than lower frequency phrasal verbs. But whether this finding also holds for collocations is an open question.

The Role of Communicative Engagement with Language in the Acquisition of Formulaic Language and Collocations

Frequency is one factor that emerges from acquisition research, but another is the facilitative value of using language for communicative purposes. Perhaps learners' engagement with language through activities like watching movies and spending time on social networking sites relates to the acquisition of L2 collocations. Most research on this question has not focused specifically on collocations, but usually on a mixed variety of formulaic language. In L1 acquisition, Nelson (1973) found that children who had referential preferences (naming things or activities and dealing with individual word items) usually learned more single words, particularly nouns. Conversely, children who had more expressive tendencies (having interactional goals; focusing on the social domain) were more likely to learn whole expressions which were not segmented. This suggests a link between the need and desire to interact and the use of formulaic sequences. This has also been demonstrated in L2 contexts. Wong-Fillmore (1976) found that formulaic sequences were relied on initially as a quick means of being communicative (albeit in a limited way) by five young Mexican children trying to integrate into an English-medium school environment. With older L2 learners, Adolphs and Durow (2004) found that the degree of social integration into the L2 community (with presumably a commensurate need to be communicative in the L2) was linked to the amount of 3-word sequences produced in the speech of L2 postgraduate students. Siyanova and Schmitt (2008) showed that spending a year in an English-speaking country (with presumably a great increase in the amount of L2 interaction) lead to better intuitions of collocation. Moreover, Schmitt and Redwood (2011) found the amount of

engagement with an L2 (extensive reading, watching films and television) differentiated higher and lower knowledge of phrasal verbs.

However, it may not be just the amount of input that is crucial, but also the quality. Siyanova and Schmitt (2007) found that the amount of exposure to native-speaking environments did not have an effect on the likelihood of using the multi-word verbs. This, however, might be explained by Adolphs and Durow's (2004) findings that socio-cultural integration was the key to their case study learner's acquisition. Similarly, Burdelski & Cook (2012) suggest that socialization can lead to the learning of formulaic language: as ideas which are important in the society are constantly stressed (e.g. politeness, honouring elders), the formulaic sequences attached to these ideas become not only frequent, but also highly salient. Bardovi-Harlig (2012) makes a similar case for pragmatics and formulaic language. Pragmatics entail using the most effective language to achieve communicative purposes, and formulaic language realizes many of these common functional objectives. Thus formulaic language should be relatively salient as it is connected with personal functional need. This all suggests that it may not be exposure per se that is important, but the kind of high-quality engagement with language that presumably occurs in a socially-integrated environment, where learners wish to use the L2 for meaningful and pleasurable communication. As Dörnyei, Durow, and Zahran (2004: 105) summarize:

> Success in the acquisition of formulaic sequences appears to be the function of the interplay of three main factors: language aptitude, motivation and sociocultural adaptation. Our study shows that if the latter is absent, only a combination of particularly high levels of the two former learner traits can compensate for this, whereas successful sociocultural adaptation can override below-average initial learner characteristics. Thus, sociocultural adaptation, or acculturation, turned out to be a central modifying factor in the learning of [formulaic language by] the international students under investigation.

Hence, we might conclude that learners who engage in greater amounts of meaningful language use will learn more formulaic sequences, especially as learners are thought to be more likely to learn sequences they find useful and meaningful for their daily lives (Ellis, 2005). This suggests that more out-of-class exposure like reading English books, watching English films/TV, and social networking in English would facilitate learning. While this seems reasonable for formulaic language in general, there is yet little evidence to demonstrate that it is also true for collocations.

In sum, there remain a number of questions regarding L2 learners' overall knowledge of collocations and what factors relate to their acquisition. Basing our research on a design intended to measure general collocation knowledge, we asked the following questions:

1. How well do Spanish learners productively know a diverse set of English collocations sampled from the COCA?
2. Do Spanish learners acquire collocations in frequency order?
3. Which method of collocation identification best relates to the collocation knowledge of Spanish learners (frequency, *t*-score, or *MI*)?
4. How do individual differences and amount of L2 instruction relate to productive collocation knowledge?
5. Does the degree of personal language use relate to productive collocation knowledge?

Methodology

Participants

The participants in this study were 108 Spanish speakers living in Spain (35 males and 73 females) with some knowledge of English as an L2. The age range was 18-64 (average 31.1). In order to study various degrees of collocation knowledge, we recruited participants from a wide range of proficiency levels, which ranged from beginner to advanced. Some of the participants were receiving formal instruction in English at the time of data collection (36), while 72 others were not. They had an L2 learning history of between 1 and 30 years (average 13.67). Only Spanish participants were chosen to control for L1 transfer effects (as evidenced by Nesselhauf, 2003, 2005).

Target Collocations

We wished to develop a sample of collocations which were as representative as possible of English collocations in general. We recognize that the nature of collocations makes it very difficult to construct a representative sample in the first place, and that no small sample can ever be truly representative of the vast domain of collocations in general. Nor is there a comprehensive list of collocations in existence to work from. Nevertheless, we feel that a sample based a primarily statistical approach (including a wide range of frequency, *t*-score, and *MI* scores) can give some indication of the wider range of possible collocations. Another reason for choosing collocations with a wide range of frequency, *t*-score, and *MI* scores is that we wished to explore how collocations selected according to these criteria relate to

learner knowledge. To begin the sampling process, we sampled 96 target collocations which varied widely along the three statistical parameters (based on the COCA), as well as meeting the following criteria:

i. All collocations had to be 2-gram, lexical collocations (e.g. *leave work*, not *do work*).
ii. The word pairs had to be considered natural English collocations by the native speakers in the piloting.
iii. Their constituent words had to be frequent (all within the most frequent 5,000 words of English, except *clockwise* (7,000)).
iv. They needed to be dispersed in their frequency, *t*-score, and *MI* rank ordering.
v. They could not be directly equivalent translations to their Spanish counterparts (e.g. *break the rules* is a direct translation of *romper las reglas*).

Materials – Collocation Test

The research instrument consisted of two sections. To obtain a measure of learner knowledge of the target collocations, a productive collocation test was developed for the first section. A productive (form recall) test format[3] was chosen to avoid guessing effects typical in multiple-choice test formats (Stewart & White, 2011). Ninety-six potential target collocations of diverse frequency were inserted into an off-line pen-and-paper productive test that took the form of a fill-in-the-gap task. After instructions in Spanish, participants were required to provide the 2-word English collocations embedded in an English sentence, which summarized or completed the information

given by the first Spanish statement. Each English sentence contained two gaps, which corresponded to each of the 2 words which formed the collocations tested. To help the participants and constrain the range of potential collocations elicited, the first letter of each of the 2 words was given.

28. Mi tía está siguiendo una dieta muy estricta porque el vestido que se compró para la boda de mi hermana le queda pequeño, y quiere entrar en él.

She wants to l_____ some w_____ by next month.

In this example, the Spanish context means "*My aunt is following a very strict diet because the dress that was bought for my sister's wedding is small, and she wants to wear it*", and so the is answer is 'lose weight'.

A pilot test with the 96 items was originally written in English and then translated into Spanish by the first author, and was checked by a second native Spanish teacher of English. A series of three pilot studies was conducted to check the validity of the test for the participants and purposes of the research. In the first stage, three native speakers of English were asked to review the item pool with only the English sentences. The aims were 1) to make sure that the English used was clear, simple, and sounded natural, and 2) to check that the English sentences did not give enough context so that the blanks could be correctly answered simply by inserting individual words which might make sense, even though the pilot participants did not know which target collocations we were prompting. At this stage, some changes were suggested in order to make the items sound more natural. Furthermore, items which

were answered easily were removed or changed in accordance with participants´ comments.

In the second phase, six Spanish speakers who had been living in the UK between 10 months and 6 years (average 1.93 years) completed the test with the English sentences alone, in the same way the English natives had done. The few successfully answered items were removed from the instrument. After that, the same six participants were asked to answer the remaining items in the near-final version of the test (i.e. with the Spanish context and then the English sentence), to ensure that the items were not ambiguous. A few additional items were removed at this stage.

In the last phase, the near-final version of the test was piloted with six participants who were similar to the target population in every way (Spanish speakers living in Spain). An item was deleted afterwards because it was found to be confusing for 5 of the 6 participants. After the piloting, the remaining well-performing items were examined, and the 50 items with the best spread of frequency, *t*-score, and *MI* scores were selected for the final test. The final target collocations (with their frequency and collocation scores) are given in Appendix 1. The statistical range of the 50 target items was: frequency (11-17,214 occurrence in the COCA), *t*-score (-2.23-130.06), and *MI* scores (0.05-45.00). The final productive test is in Appendix 2.

Materials – Questionnaire

The second section of the instrument was a questionnaire designed to collect information about the degree to which participants engaged with and used the L2. It began with items about the participants' individual differences (gender, age, and

proficiency). We then explored the amount of input learners had received in an instructed context, asking about the number of years of each participant had studied English. We also asked about the amount of input which learners had received through their personal weekly use of English outside the classroom. This 'language use' factor was made up of several personal types of usage. The first type of language use was reading in English. Reading facilitates the learning of individual words (e.g. Horst, Cobb, & Meara, 1998), and this may be true of formulaic language as well, as Schmitt and Redwood (2011) found that the amount of reading related to knowledge of phrasal verbs. Input can also come in the form of audio/video input, and so our second type of language usage was watching English films, video, or TV, and the third type was listening to English language music.

Social networks, such as Twitter, Facebook, Skype or MySpace have recently grown extremely popular in all developed countries, and people of all ages use them daily to communicate and socialize with others. English is often the lingua franca of this type of communication, and so for the fourth type of language activity, we were interested in discovering whether Spanish participants use social networking to communicate in English and, if so, how often, with the purpose of evaluating its influence in the acquisition of collocations. For each of the above four personal language use activities, we asked participants to indicate how many hours per week they participated in the activity: 0-1, 1-2, or more than 2 hours a week.

The most intensive type of language use usually comes with immersion in English-speaking countries. It is widely recognized that the immersion in an L2 environment improves general language learning and facilitates the process (Pawley and Syder, 1983; Cummins, 1998; Freed *et al*. 2004). It has also been proposed that formulaic sequences are so closely related to everyday target language that cannot be learnt

most efficiently unless the learner is immersed in the L2 culture and involved in the life of the L2 community (Dörnyei *et al*. 2004:87). As a fifth type of language use, we asked participants whether they had spent three months or more in English-speaking countries. See Appendix 3 for the complete language background and use questionnaire.

Procedure

The administration of the test was carried out in Spain. No time limits were set, and the completion time ranged from 25 to 50 minutes (average 42.5). The test was administered either individually or in small groups, and all participants were offered the version of the test they felt more comfortable with, the Spanish or the English. The tests were conducted face to face (including the few that were administered via Skype) and supervised by the first author. Instructions for completion were provided, and participants made aware of the confidentiality and voluntary nature of the test, as well as the general purpose of the study.

The language use variables were marked as follows: participants who engaged in the language use activity 0-1 hours were coded as 1, those who engaged 1-2 hours coded as 2, and those who engaged more than 2 hours coded as 3. Three everyday activity types (reading, films/TV and social media) were also merged into a composite 'exposure to language' variable. The ordinal scale results from each of these variables were correlated with the collocation knowledge scores with a Kendall's tau correlation statistic.

Results and Discussion

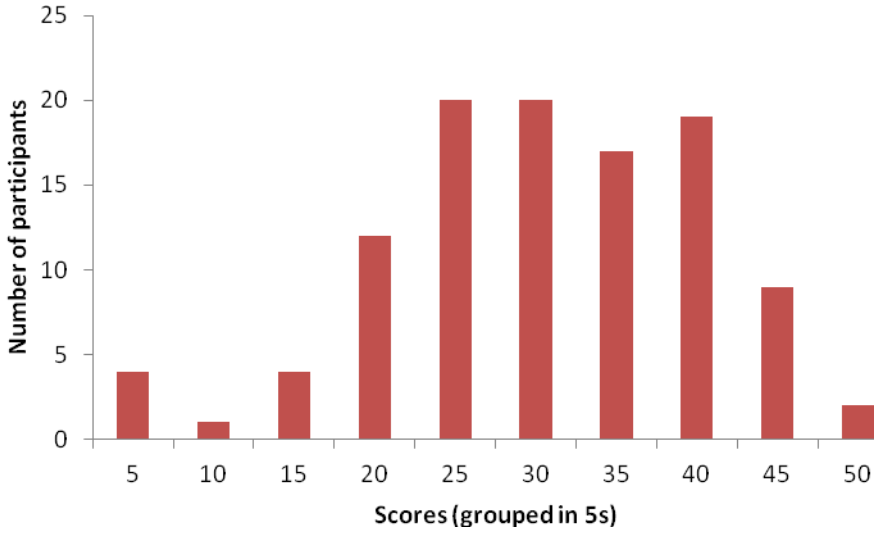
To What Degree Do Spanish Speakers Know Collocations?

It is worth noting that our productive test format was not amenable to either guessing or translation from the L1 Spanish. Also, target items included a number of quite low-frequency collocations: *clockwise direction* – only 33 instances in the 450-million word COCA, *overcome* (a) *difficulty* – 25 instances, and *exploit resources* – 11 instances. With this in mind, Table 1 shows that the mean score is 28.29 out of a possible 50 (56.6%). This shows that many of the Spanish participants knew a substantial number of the target collocations, although the large standard deviation shows that there was considerable variation across the sample. This variation is also shown by the range, with the number of correct answers varying from a minimum of 1 to a maximum of 46 out of 50. The participant performance is graphically illustrated in Figure 1, which shows that very few participants scored less than 15 (9 people) or more than 40 (11 people), with the vast majority (70%) scoring between 21 and 40 (76 people). Clearly, this Spanish group of participants have considerable collocation knowledge, and are able to produce the written form of a substantial number of collocations, at least as indicated by this test format. However, it is a matter of speculation the degree to which the participants would be able to employ this knowledge in their own free writing and speaking.

Table 1 Descriptive statistics of the participants' test performance (N=108)

	M	SD	Min	Max
Participant scores (Max=50)	28.29	9.74	1	46

Figure 1 Distribution of collocation test scores



Which Method of Collocation Identification Best Relates to Collocation Knowledge?

Our Spanish learners acquired collocations to a substantial degree. We next investigated which statistical method of collocation identification best relates to this knowledge. In particular, we explored corpus frequency, *t*-score and *MI*. That is, does L2 acquisition of collocations mainly relate to those collocations' frequency of occurrence (raw frequency or *t*-score) or their strength of association (*MI*)?

The percentage of correct answers by our Spanish participants for each collocation was correlated with each of the three collocation measures (Table 2). Raw corpus frequency correlated with the collocation knowledge of participants at .45 with an r_s^2

of just over 20%. *T*-score was just less than this, with a correlation of .41 and r_s^2 of just below 17%. The similar results are not surprising, as *t*-score is heavily weighted towards frequency in its calculations. This means raw frequency was related slightly more strongly to the learners' knowledge of the target collocations than *t*-score was, and has the advantage of not requiring a calculation. Conversely, *MI* score did not show any significant relationship with collocation knowledge. This indicates that increasing the 'tightness' of the combinational bonding does not seem related to collocation learning. Rather, in line with Durrant and Schmitt (2009) and Ellis, *et al*. (2008), it seems that the frequency of the collocation as a whole is the more important factor for second language learning of collocations, although our results show that the relationship is only a moderate one with 20% co-variance.

Table 2 Correlations between knowledge of collocations and three methods of collocation identification

	Raw Frequency[a]		*t*-score		*MI*	
	Correlation	r^2	Correlation	r^2	Correlation	r^2
Participants' scores	.45**	20.3[b]	.41**	16.8	-.16	2.6

a. Frequency from the COCA
b. r^2 reported in percentage
** Spearman: $p < .01$

As corpus frequency was the variable that best related to acquisition, it is useful to further explore what the 20% covariance tells us. The relationship between collocation knowledge and frequency can be best appreciated graphically. Figure 3 shows the correspondence using a moving average of five collocations to reduce the

effect of individual item variation. This can be compared to the apparently random (non)relationship between collocation knowledge and *MI* score (Figure 4).

Figure 3 Knowledge and frequency of collocations (moving average of 5)

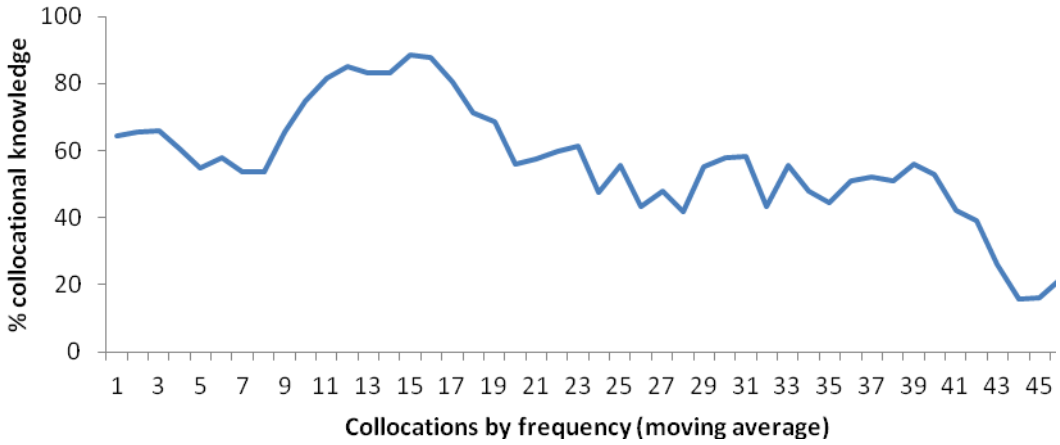
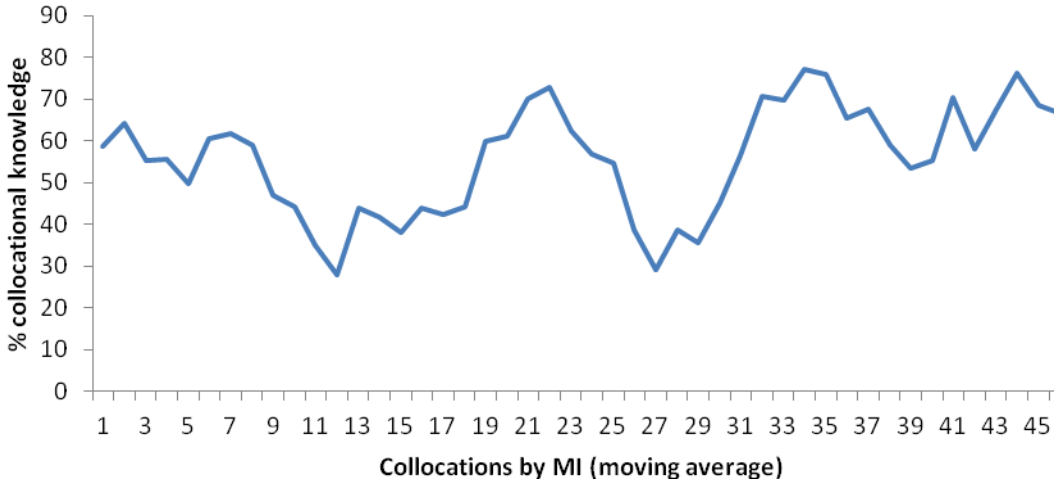


Figure 4 Knowledge and *MI* score (moving average of 5)



It is clear from the curve in Figure 3 that collocation knowledge is not strongly dependent on the frequency of collocations, although it does seem to have some

influence. This is quite different to the knowledge of individual words, where frequency has been shown to be a strong determining factor (e.g. Schmitt, Schmitt, and Clapham, 2001). The figure shows a jagged, but slightly downward trend, with a noticeable trailing off of knowledge for the least frequent collocations, and a cluster of relatively well-known collocations at the 10-20 frequency rank order. Overall, we cannot say that collocations are learned in frequency order, and that the highest frequency collocations are learnt better, but it seems clear that very infrequent collocations are learnt worse.

The above reports whole-group findings, but such summative descriptions often hide interesting information. Looking beyond the whole group scores, we wondered whether the behavior of learners with relatively better collocation knowledge was similar to or different from learners with relatively weaker knowledge. In order to explore this, we divided the participants in thee groups according to their total collocation scores: low (1-20, *N*=21), medium (25-35, *N*=40), and high (40+, *N*=14).[4] Given the results from Figure 3, we might expect that the high group would know most of the high- and mid-frequency collocations, and would only tail off at the lowest-frequency ones. We might also expect that the low learners would only know the highest-frequency collocations, and very few others. The mid group would be somewhere in between, and might be expected to know the highest frequency collocations best, with a tail-off at the lowest frequency ones. These hypothetical expectations might be visualized as something like Figure 5 (ordered by frequency, from high to low).

The actual results in Figure 6 show that frequency does not have a similar effect on the three groups. The high group behaves roughly as expected: the vast majority of high proficiency group knew almost all of the 40 highest-frequency collocations, but this tailed off badly over the last 10 least-frequent ones. Thus, frequency certainly seems to explain this knowledge profile. The medium- and low-knowledge groups have a similar tailing off, but in contrast also show a great deal of variation across the profile. Interestingly, the medium and low groups have nearly identical curves. If a collocation was relatively easy or difficult for the mid-proficiency group, it was as well for the low-frequency group, and vice versa. The undulating profiles suggest that the knowledge of high- and mid-frequency collocations do not depend greatly on the corpus frequency of those collocations for the learners in these groups.

These results indicate that although frequency seems to have some relationship with the acquisition of collocations, this effect is slight and cannot be used as the major predictor of collocation learning, but only as one factor of influence. This is in contrast to Ellis *et al*. (2008), who found that frequency of academic formulas is a good predictor for L2 learners´ processing.

Figure 5 Hypothetical knowledge profile of high-, mid-, and low-knowledge groups according to frequency order

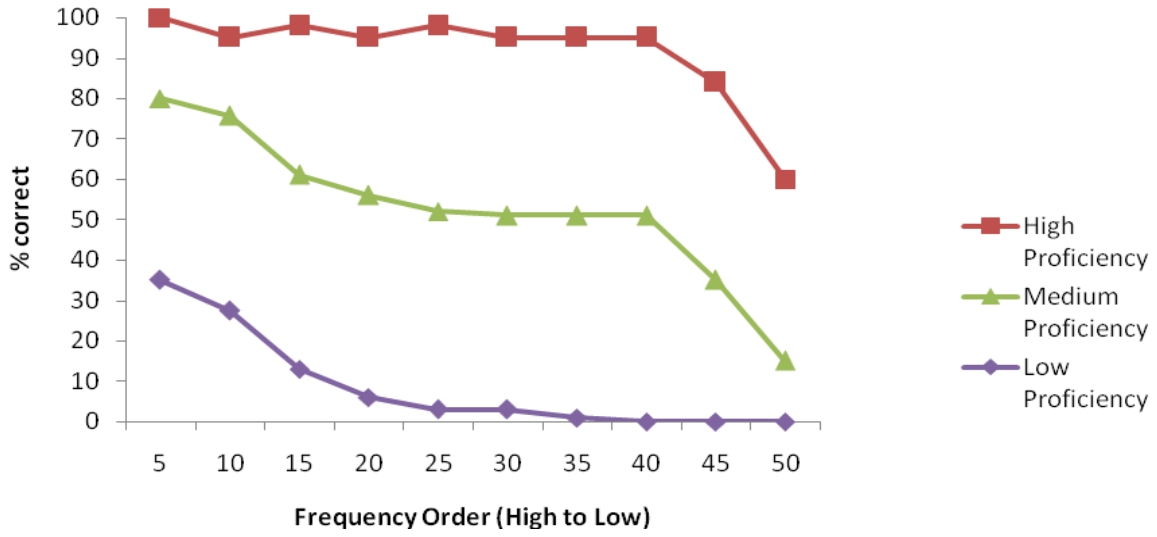
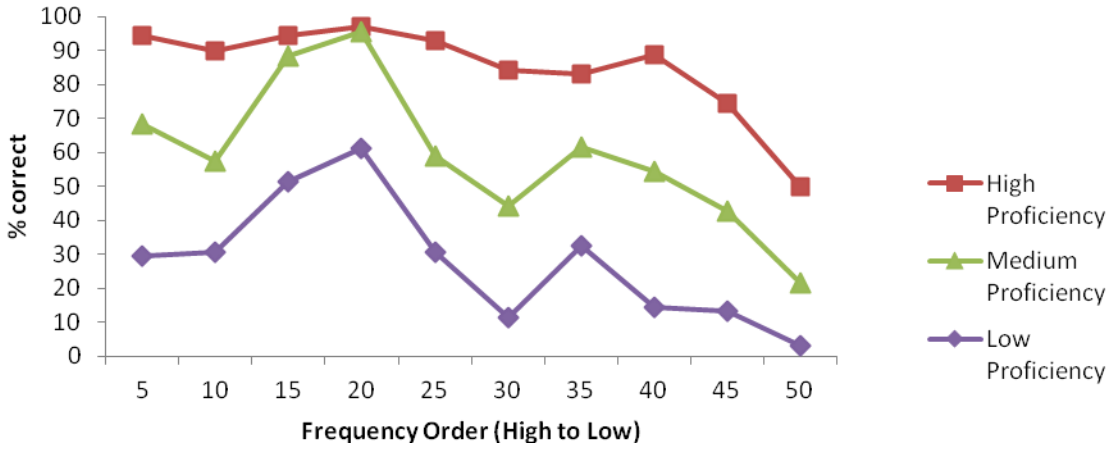


Figure 6 Actual knowledge profile of high-, mid-, and low-knowledge groups according to frequency order



Does Individual Differences and Language Study Relate to the Knowledge of Collocations?

If the combinationality of collocations (*MI*) does not relate to collocation knowledge, and frequency does so only to a limited degree, what does relate to collocation knowledge? In order to answer this question we examined the relationships of a number of factors with the collocation knowledge of our participants. We first looked at the individual difference factors of gender, age, and proficiency. There was no effect for gender (Male vs. Female: t (106) = -.40, p =.70) and age had only a very weak relationship (.20, r^2 = 4.0, Pearson: p<.05). We asked participants to self-rate their L2 proficiency according to a three-part categorization: beginner, intermediate, or advanced, and the correlation between this self-rating and collocation knowledge was strong: .73, r^2=53.3, Kendall's tau: p<.001. However, as proficiency self-ratings can be rather subjective, this result needs to be treated with caution.

It has been suggested that collocations can be taught, although most textbooks do not give them much time or attention (Brown, 2011). If so, one would expect that the amount of study and formal instruction would have an effect on collocation learning. This proved to be the case here, as the correlation between the years of study that the participants had engaged in and the correct answers they gave to the test was .45, r^2 = 20.3, Pearson: p<.001. Therefore, the results indicate that the amount of language study and instruction related to collocation knowledge, although rather moderately. Overall, these results are in line with those of Schmitt *et al*. (2004) who found that instruction appears to facilitate the acquisition of formulaic sequences.

Does Use of Language Relate to the Acquisition of Collocations?

It has been suggested that the amount and type of informal exposure learners have to English outside the classroom can affect the degree to which collocations are learned. We asked the participants about their personal engagement with English through a number of activities they might carry out in their daily lives: reading, watching TV or films, listening to music, using social media, and visiting an English-speaking country. Correlation analysis showed that all of these variables did indeed relate to collocation knowledge, with the exception of listening to music (Table 3): immersion .64, reading .61, TV/films .38, social media .33, and listening to music *ns*. Some of these correlations were higher than the .45 we found for corpus frequency. Combining the everyday exposure variables of reading, TV/films, and social media,[5] we find that composite informal exposure to the target language correlates at .56, explaining over 31% of the collocational knowledge that participants showed, meaning that the more language input learners receive, the more likely they are to learn more collocations.

Table 3 Correlations between language use and knowledge of collocations

	Knowledge of collocations	
	Correlation	r^2
Reading	.61**	37.2[a]
Watching films, video, or TV	.38**	14.4 [a]
Listening to music	.14	1.96 [a]
Social networking	.33**	10.9 [a]
Immersion in English-speaking countries	.64***	41.0 [a]
Composite exposure to English[b]	.56**	31.4 [a]

** Kendall's tau: $p < .001$
*** Biserial: $p < .001$
a. r^2 reported in percentage
b. Composite score includes Reading, Watching films/TV, and Social networking

General Discussion

Our results indicate that our Spanish participants did know a substantial percentage of collocations, as evidenced by an average of over 56% correct answers on a prompted productive collocation knowledge test. Furthermore, these figures are likely to be a good representation of their knowledge, as the answers could not easily be the result of guessing or L1 transfer. This result runs counter to the assertion of some researchers who have claimed that the knowledge that NNSs have of collocations is low (Bahns and Eldaw, 1993; Farghal and Obeidat, 1995; Laufer and

Waldman, 2011). However, it is interesting to consider whether 56% should be considered 'good' knowledge or not. On one hand, the collocation sample included a range of collocations including some of quite low frequency (e.g. *exploit resources*: 11 instances in the 450-million word COCA; *clockwise direction*: 33 instances). This would suggest the participants' performance is relatively strong. Conversely, the mean number of years which participants reported studying English was 13.67 years. From this point of view, one might expect much higher scores, and would thus conclude that collocations are indeed difficult to pick up from language study. Ultimately, a reasonable conclusion seems to be that our Spanish learners did demonstrate knowledge of a sizeable number of collocations, although falling short of what might be expected from native, or very proficient nonnative, speakers.

An interesting research direction for the future would be to explore how the level of collocation knowledge required by our test relates to the ability to use English skillfully and appropriately in the four skills. Laufer and Waldman's (2011) very low collocation results when analyzing Israeli student compositions suggest that it might be much easier to produce collocations when prompted on a test like ours, than to freely produce collocations in one's own writing, and that there is a big difference between *knowing* a collocation and being able to *use* it. However, this remains speculation, because although Laufer and Waldman's study included a range of L2 proficiency levels (just as ours did with Spanish participants), it is impossible to know how the participant proficiency levels of the two studies compare.

The study has also revealed that, unlike for the acquisition of individual words, corpus frequency is not as strong a factor as might have been expected for

collocation knowledge, although it is still a better predictor than the *MI* or *t*-score. It showed that even though corpus frequency seems to relate to some degree to the acquisition of collocations, it only explains just over 20% of the collocation knowledge tested here, mainly indicating that the lowest-frequency collocations were poorly known.

Furthermore, the findings of this study are in line with a recent study by Schmitt and Redwood (2011), which showed the same moderate effect of frequency on the acquisition of phrasal verbs. Taken together, it seems that frequency of occurrence is not an adequate predictor of formulaic language acquisition. This leads us to believe that the acquisition of formulaic language (at least collocations and phrasal verbs) relies on more than just frequency of exposure, or at least frequency as derived from a corpus of native English. The results from our usage questionnaire suggest that engagement with collocations in everyday communicative situations (reading, watching TV/films, social networking, and immersion) may well be just as important a factor. As might be expected, immersion had a high correlation (.64), but even the everyday composite correlation was higher (.56) than the correlation for corpus frequency (.45). Therefore, if collocation learning is facilitated by everyday use in communication, teachers and material designers would do well to incorporate these types of activities into their syllabuses and materials as much as possible.

Overall, we propose that the results of this study have the following implications for language teaching. First, we found that our L2 learners knew a substantial percentage of the collocations tested. Therefore, these findings suggest that there is hope for collocation learning, and that collocations (and formulaic language in general) should stop being considered a subject which is overly challenging to learn

in language instruction. The amount of study was shown to have a positive relationship (r^2=20%), even though language instruction in Spain seldom highlights the notion of collocation. Moreover, research focusing on explicit instruction of collocations (e.g. Laufer & Girsai, 2008; Peters, 2014) shows that it can be effective, especially if there is enough learner engagement and repetition. Even if the instruction only improves general language proficiency, our self-report data suggests that proficiency itself seems to relate strongly to collocation knowledge (r^2=53%).

Second, a large number of our participants knew a sizeable number of the target collocations. This suggests in any class group, there will be a considerable number of collocations known by some students, but not by all. The implication is that group activities might be useful where students do tasks designed to elicit and use the combined collocation resources of the group. Newton (1995) found that learners could learn new words which were used by other members of their group in the process of completing group tasks. This approach may well also work for collocation learning.

Third, usage-based theories posit that acquisition is essentially linked to the amount of language exposure (e.g. Ellis, 2002). Extensive reading is often recommended as a way of maximizing this exposure outside the classroom (e.g. Day and Bamford, 1998). This study indicates that consistent reading is also useful for collocation acquisition (r^2=37%). But our results show that there are many other kinds of exposure that are also useful beyond reading. Teachers should strongly encourage their students to take advantage of any English-exposure opportunities that are available. Our research demonstrated that watching English TV/films and using social networking sites were useful. Indeed, students may well be doing these things already, but if not, teachers could promote their use by activities such as making

worksheets based on movies, or setting up ‘penpal-type’ relationships on social networking sites.

Fourth, corpus frequency (as indicated by the COCA) relates to collocation acquisition, but only moderately. However, our participants knew the target collocations better if they used English in everyday situations. This is congruent with Ellis´s (2001) claim that frequent collocations which fulfill a useful and meaningful communicative function will be more salient to learners, and therefore more likely to be learnt than those with less useful functions. Likewise, Slobin (1997) claimed that there are many other determinant factors in the acquisition of language other than frequency, like semantic basicness, salience, communicative intent or relevance. This suggests that L2 learners might better know those collocations which are likely to be encountered in daily situations, and therefore to have a more useful communicative function, compared to collocations whose function is more restricted to specific contexts, such as the most ‘academic’ collocations. For example, 99 participants knew the collocation *social networks* (frequency 899) because, independently of its frequency, it is an everyday reality everyone is familiar and involved with, while only 50 knew *human being* (frequency 5,737). Therefore, frequency based on general English native speaker corpora may not be the best way of sequencing collocations in instruction, as it may not reflect actual learner exposure very well. This makes us wonder whether frequency as derived from specialized corpora better representing learner usage might be a superior way to predict collocation knowledge. For example, future research could derive frequency figures from corpora of social networking language or films, and explore whether these figures align more closely with L2 learner collocation knowledge. If so, then these

figures might prove useful as a means of selecting collocations in a principled manner for future teaching materials.

Finally, all of the findings above must be interpreted in light of the inevitable limitations of this study. We used a statistical approach to collocation identification, and a sample of collocations identified with a different method (e.g. a phraseological approach) might be known to a greater or lesser degree. Our study is also limited by only using 50 collocations to represent the vast range possible, and only having Spanish participants. Lastly, our c-test format shows whether collocations could be produced upon prompting, but we do not know whether participants could produce them in their free writing or speaking. Despite these limitations, we feel that our study still provides useful initial insights into the amount of collocation knowledge that L2 learners might amass, and the role of frequency and input into that acquisition.

Conclusion

Our study shows that L2 learners typically know a substantial number of collocations, providing some evidence to counteract the notion that collocations are too hard for learners. Given the importance of collocations for accurate and appropriate language use, this is good news, and will hopefully encourage both researchers and practitioners to renew their interest in how to best facilitate collocation learning. The approach will almost certainly need to include a component which encourages learners to engage with English in their everyday language-based activities.

Notes

1. In contrast, psycholinguistics has tended to investigate idioms (Siyanova-Chanturia, 2013).

2. Although Martinez and Murphy (2011) worked with multiword expressions rather than collocations in particular, many of their instrument items appear to be collocations.

3. Webb (2008) argues that c-test formats which provide the initial letters of target words, although used to measure productive vocabulary knowledge, may in fact measure receptive knowledge to some degree.

4. In order to ensure the groups were clearly distinct form each other, we deleted participants with scores between 21 and 24 collocations (17 people) and between 36 and 39 (16 people).

5. Immersion was a dichotomous variable, and listening to music was nonsignificant.

References

Adolphs, S. & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 106-126). Amsterdam: Benjamins.

Annual Review of Applied Linguistics. (2012). Vol. 32: Formulaic language.

Bahns, J. & Eldaw, M. (1993). Should We Teach EFL Students Collocations? *System*, 21,101-114.

Bardovi-Harlig, K. (2012). Formulas, routines, and conventional expressions in pragmatics research. *Annual Review of Applied Linguistics, 32,* 206-227.

Barfield, A. and Gyllstad, H. (Eds.). (2009). *Researching collocations in another language: Multiple interpretations.* Basingstoke: Palgrave Macmillan.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research 10*, 245–261.

Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research, 15,* 83-97.

Burdelski, M. & Cook, H.M. (2012). Formulaic language in language socialization. *Annual Review of Applied Linguistics, 32,* 173-188.

Bybee, J.L. & Hopper, P. (2001). *Frequency and the emergence of language* structure. Amsterdam: John Benjamins.

Cowie, A. P. (Ed.). (1998). Phraseology: theory, analysis, and applications. Oxford: Oxford University Press.

Cummins, J. (1998). Immersion education for the millennium: What we have learned from 30 years of research on second language immersion. In M.R. Childs and R.M. Bostwick (Eds.), Learning through two languages: research and practice. Second Katoh Gakuen International Symposium on Immersion and bilingual education (pp. 34-47). Japan: Katoh Gakuen.

Davies, M. (2008-) The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/. [Accessed from June to August, 2013].

Day, R. R. & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge: Cambridge University Press.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair and M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp.51-68). Amsterdam: Rodopi.

Dörnyei, Z., Durow, V., & Zahran, K. (2004). Individual differences and their effects on formulaic sequence acquisition. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 87–106). Amsterdam: John Benjamins.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics* 47, 157–177.

Durrant, P. & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research,* 26, 163-188.

Ellis, N.C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction*. Cambridge: Cambridge University Press.

Ellis, N.C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.

Ellis, N.C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition,* 27, 305–352.

Ellis, N.C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics and TESOL. *TESOL Quarterly*, 42, 375-396.

Farghal, M. & Obiedat, H. (1995). Collocations: A Neglected Variable in EFL. *IRAL*, 33, 315-331.

Freed, B.F., Segalowitz, N. & Dewey, D.P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, 26, 275-301.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 79–100). Oxford: Oxford University Press.

Granger S., Paquot M. & Rayson P. (2006). Extraction of multi-word units from EFL and native English corpora: The phraseology of the verb 'make'. In Häcki Buhofer, A. and H. Burger (eds.), *Phraseology in Motion I: Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie* (Basel, 2004). Baltmannsweiler: Schneider Verlag Hohengehren. pp. 57-68.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*, 237–260.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*, 207–223.

Howarth, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161–186). Oxford: Oxford University Press.

Irujo, S. (1993). Steering clear: Avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching 31*, 205-219.

Laufer, B. & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*, 1-23.

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second-language writing: A corpus analysis of learners´ English. *Language Learning*, 61, 647-672.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. Longman.

Lorenz, G. (1999). *Adjective intensification – learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

Martinez, R. & V. Murphy. (2011). Effect of frequency and idiomaticity in second language reading comprehension. *TESOL Quarterly 45*, 267-290.

Martinez, R. & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics* 33, 299-320.

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics, 32*, 111-129.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics* 32, 129–48.

Nation, I.S.P. (2001). *Learning Vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge University Press.

Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, Serial no. 149, nos 1-2.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24, 223–242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Newton, J. (1995). Task-based interaction and incidental vocabulary learning: A case study. *Second Language Research 11*, 159-177.

Pawley, A. & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (Eds.), *Language and communication* (pp.191-226). London: Longman.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *Language Learning Journal, 40*, 65-79.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research, 18*, 75-94.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2004) (Ed.). *Formulaic sequences: Acquisition, processing and use.* Amsterdam: Benjamins.

Schmitt, N. (2010). *Researching vocabulary: a vocabulary research manual*. Basingstoke: Palgrave McMillan.

Schmitt, N., Dörnyei, Z., Adolphs, S., and Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: John Benjamins Press. pp. 55-86.

Schmitt, N. & Redwood, S. (2011). Learner knowledge of phrasal verbs: a corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin, and M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp.173-209). Amsterdam: John Benjamins.

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing 18*, 1: 55-88.

Sinclair, J. (2004). *Trust the text: Lexis, corpus, discourse*. London: Routledge.

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research. *Mental Lexicon, 8,* 245-268.

Siyanova, A. & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *International Review of Applied Linguistics, 45*, 119-139.

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*, 429-458.

Slobin, D.I. (1997). The origins of grammaticizable notions: Beyond the individual mind. In: D.I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 5* (pp. 265–323). Mahwah, NJ: Erlbaum.

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly, 45*, 370-380.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics 28*, 46-65.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*, 79-95.

Webb, S., Newton, J. & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*, 91-120.

Wong-Fillmore, L. (1976). *The second time around*. Doctoral dissertation: Stanford University.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Appendix 1 List of target collocations (with frequency, t-score and MI information)

	Collocation	Raw frequency	Frequency of individual words[a]	T-score	MI
1	Last night	17,214	Last: 357,284 Night: 194,531	130.06	6.85
2	Long ago	8,455	Long: 277,641 Ago: 143,237	91.02	6.62
3	Nuclear weapons	7,364	Nuclear: 41,511 Weapons: 36,130	85.78	11.15
4	Interest rates	7,026	Interest: 76,183 Rates: 43,539	83.74	9.94
5	Human being	5,737	Human: 123,875 Being: 309,533	74.65	6.12
6	To spend time	4,875	Spend: 154,443 (lemma) Time: 735,882	66.31	16.98
7	Hard work	4,763	Hard: 136,530 Work: 361,432	67.47	5.48
8	Wide range	4,653	Wide: 42,179 Range: 50,815	68.15	9.98
9	Vast majority	4,421	Vast: 20,266 Majority: 41,255	66.46	11.26
10	To commit suicide	2,500	Commit: 31,154 (lemma) Suicide: 16,507	49.98	45.00
11	Human nature	2,494	Human: 123,875	49.59	7.15

			Nature: 65,802		
12	To lose weight	2,438	Lose: 170,401 (lemma) Weight: 44,812	49.04	28.63
13	Eye contact	1,795	Eye: 48,659 Contact: 35,642	42.28	8.91
14	To hold hands	1,596	Hold: 217,985 (lemma) Hands: 100,440	38.77	17.15
15	To find (a) job	1,499	Find: 463,482 (lemma) Job: 128,885	35.39	9.83
16	Free time	1,422	Free: 101,409 Time: 735,882	33.44	3.14
17	Illegal immigrant	1,344	Illegal: 20,093 Immigrant: 8,757	36.65	9.71
18	Live music	1,022	Live: 115,646 Music: 117,880	31.05	5.12
19	To keep quiet	1,000	Keep: 271,808 (lemma) Quiet: 30,401	31.06	21.22
20	Social networks	899	Social: 156,256 Networks: 15,241	29.81	7.45
21	Positive attitude	751	Positive: 45,730 Attitude: 20,812	27.33	8.52
22	Early retirement	710	Early: 141,907 Retirement: 15,561	26.47	7.22
23	Financial problems	623	Financial: 58,302 Problems: 108,249	24.42	5.52
24	Annual income	577	Annual: 33,165 Income: 35,891	23.91	7.81
25	Volunteer work	564	Volunteer: 10,025 Work: 361,432	23.42	6.18

26	To resist (the) temptation	525	Resist: 15,173 (lemma) Temptation: 3,177	22.91	43.25
27	Married couple	516	Married: 45,037 Couple: 75,930	22.39	6.13
28	To raise awareness	494	Raise: 107,488 (lemma) Awareness: 14,193	22.08	27.68
29	Physical contact	429	Physical: 60,312 Contact: 35,642	20.49	6.53
30	Night shift	408	Night: 194,531 Shift: 22,161	19.74	5.46
31	Traffic jam	379	Traffic: 24,283 Jam: 4,664	19.46	10.60
32	To face (a) challenge	349	Face: 226,185 (lemma) Challenge: 38,756	17.67	15.66
33	Organic food	282	Organic: 12,775 Food: 109,622	16.61	6.55
34	To leave work	251	Leave: 367,675 (lemma) Work: 361,432	-2.23	0.05
35	Day trip	235	Day: 347,473 Trip: 38,097	13.47	3.04
36	Controversial issue	220	Controversial: 11,405 Issue: 100,607	14.67	6.48
37	To reveal (a) secret	195	Reveal: 51,361 (lemma) Secret: 34,111	13.69	18.85
38	Personal belongings	187	Personal: 76,997 Belongings: 1,946	13.65	9.18
39	Unlimited access	141	Unlimited: 3,325 Access: 46,752	11.85	8.75
40	To prescribe (the)	133	Prescribe: 6,986	11.52	38.26

	medication		(lemma) Medication: 7,137		
41	To withdraw money	102	Withdraw: 12,807(lemma) Money: 188,742	9.58	18.99
42	Foreign accent	101	Foreign: 64,606 Accent: 6,197	9.96	6.87
43	To live abroad	100	Live: 337,477 (lemma) Abroad: 10,140	9.26	18.77
44	To destroy evidence	81	Destroy: 30,080 (lemma) Evidence: 75,526	8.46	12.53
45	Entry requirements	75	Entry: 12,409 Requirements: 15,876	8.61	7.46
46	To generate jobs	55	Generate: 27,340(lemma) Jobs: 53,170	6.99	16.85
47	Unforeseen circumstances	46	Unforeseen: 803 Circumstances: 20,239	6.78	10.36
48	Clockwise direction	33	Clockwise: 2,084 Direction: 34,734	5.72	7.73
49	To overcome (a) difficulty	25	Overcome: 13,968 (lemma) Difficulty: 14,478	4.91	9.93
50	To exploit resources	11	Exploit: 8,477 (lemma) Resources: 48,982	3.05	8.64

a. In cases where the collocation includes a verb, the frequency figures are given for lemmas instead of word families.

Appendix 2 The productive collocation test

Estamos llevando a cabo un estudio sobre el conocimiento de vocabulario y expresiones en inglés que tienen los españoles. Para ayudarnos en nuestra investigación, por favor, completa el siguiente test. Tienes un contexto escrito en español y una frase en inglés que resume o completa la información previa. Completa los huecos con información dada o implícita en el contexto. Solo necesitas 2 palabras para cada pregunta (una palabra en cada hueco), y, para ayudarte, la primera letra de cada palabra ya viene dada.

*El test es **absolutamente voluntario y confidencial**, y los resultados serán utilizados **solo** con propósitos de investigación. Sin embargo, necesitamos que firmes una **hoja de consentimiento** que encontrarás al final de este test. ¡Apreciamos mucho tu participación!*

Hay 50 preguntas. Intenta contestar tantas como puedas (cuantas más mejor), pero no te preocupes si no las puedes contestar todas.

¡Buena suerte!

[We are carrying out a study on Spanish people's knowledge of English vocabulary and expressions. To help us in our research please complete this test. You will find a context written in Spanish and then a sentence in English which summarizes or completes that information. Complete the slots with information given or implicit in the context. You will only need 2 words for each question (one word in each blank), and the first letter of each word is shown to help you.

The test is **volunteer and completely confidential**, and the results will be used **only** for research purpose. However, you still need to sign a **consent form** at the end of the test. Your participation is much appreciated!

There are 50 questions. Try to answer as many questions as possible (the more, the merrier), but do not worry if you cannot answer them all.

Good luck!]

1.	*Cuando tenía 12 años me gustaba mucho jugar con muñecas, pero eso fue hace ya mucho tiempo.* L_____ a_____ I liked playing with all kind of dolls.
2.	*Mi tío Gerarld es guarda de seguridad en un centro comercial, y trabaja a turnos por semanas. Esta semana ha estado de día, así que la siguiente le toca de noche, lo cual no le agrada.* He is not very happy because he has to do the n_____ s_____ next week.
3.	*En trabajos que exigen mucho esfuerzo físico, como la minería, es usual que los trabajadores se jubilen antes de tiempo, pudiendo hacerlo desde los 48 años.*

	They can apply for an e_____ r_____ at the age of 48, especially if they suffer from a work-related illness.
4.	*Ahora muchos supermercados ofrecen comida "bio", que no ha sido tratada con productos químicos de ninguna clase, pero son más caros que los normales.* Many families still can´t afford o_____ f_____ because it is more expensive.
5.	*Hoy en día muere mucha más gente por causas naturales que por otras causas.* The v_____ m_____ of people die of natural causes.
6.	*"Perdón por llegar tan tarde al trabajo, pero había un atasco enorme y estuve parado durante más de tres horas".* He was in a t_____ j_____ for more than 3 hours.
7.	*Dar dos besos a la hora de saludar a alguien es propio de la cultura hispana. Sin embargo, en muchas culturas la gente no se toca cuando se saluda.* Ph_____ c_____ when greeting is uncommon and inappropriate in many cultures.
8.	*En el Reino Unido es muy común entrar en un pub y encontrarse con músicos tocando en directo.* Most of the pubs in the UK have free l_____ m_____.
9.	*Quiero empezar a comer menos y más sano, pero cuando tengo hambre no me puedo resistir a comer patatas fritas.* I just can´t r_____ the t_____ to eat unhealthy snacks.
10.	*Pedro ha decidido enfrentarse al reto de estudiar una carrera sin dejar de trabajar a tiempo completo, y nosotros le vamos a ayudar y apoyar.* Pedro is not going to f_____ that ch_____ alone.
11.	*En una entrevista de trabajo recuerda que es muy importante mantener siempre el contacto visual con tu entrevistador.* Making e_____ c_____ with the interviewer may help you to get the job.
12.	*Mi madre normalmente entra a trabajar a las 8:00 am, y sale a las 5:30 pm.* Now it is 5:15 pm so she will l_____ w_____ in 15 minutes.
13.	*A pesar de los esfuerzos de las autoridades sanitarias por despertar conciencia acerca de los peligros del tabaco todavía es mucha la gente que fuma.* Health authorities want to r_____ a_____ about the dangers of smoking.

14.	*Hoy tuvieron que cerrar la biblioteca debido a ciertos imprevistos, y van a mantenerla cerrada durante una semana.* U_____ c_____ forced the library to close.
15.	*Cuando los tipos o tasas de interés suben, las personas que tienen deudas tienen que pagar más.* Therefore, most people are happy when the i_____ r_____ are low.
16.	*Los distintos países intentan aprovechar al máximo todos los recursos que les ofrece la naturaleza de esa zona, como agua, carbón, madera, petróleo…* They want to e_____ all the r_____ available in the area.
17.	*Como Paco no encontraba trabajo y estaba ya cansado de estar desempleado, decidió trabajar como voluntario en distintas organizaciones.* Doing some v_____ w_____ gave him great satisfaction.
18.	*Mi amiga Rebeca quería estudiar un máster en Noruega, pero no pudo porque no cumplía los requisitos de acceso necesarios.* She didn't satisfy the e_____ r_____ so she couldn't do it.
19.	*Mi vecina heredó la empresa de su marido, pero resultó tener más deudas que ganancias y ahora está atravesando una mala situación económica.* She is now facing serious f_____ p_____.
20.	*Si te registras como usuario de esta página web tienes acceso a todos los contenidos que ofrece.* Only registered users can have u_____ a_____ to all of the content.
21.	*Desde que comenzase la crisis en 2008 ya no hay tantos inmigrantes sin documentar que intenten entrar en el país.* Thus, the total number of i_____ i_____ in the country has decreased.
22.	*En esta tienda no aceptamos pagos con tarjeta, pero hay un cajero automático al cruzar la calle.* You will need to go there and w_____ some m_____.
23.	*Si crees en ti mismo y eres positivo puedes superar todas las dificultades que se te presenten.* It is easier to o_____ any d_____ if you are confident.
24.	*Enfermedades como la depresión están fuertemente ligadas al suicidio.* Therefore, if people suffer from depression they are more likely to c_____

	s_____.
25.	*Mi hermano y su novia nunca van de la mano delante de mis padres.* They think it is inappropriate to h_____ h_____ in front of his parents.
26.	*La abuela de mi mejor amiga tiene una receta secreta para hacer tarta de queso que solo ella conoce, y dice que se la llevará a la tumba.* She says she will never r_____ her s_____ recipe to anybody.
27.	*Kate se pasa todo el día trabajando, y nunca encuentra el momento para hacer deporte.* Hence, because Kate s_____ all her t_____ working she is not really fit.
28.	*Mi tía está siguiendo una dieta muy estricta porque el vestido que se compró para la boda de mi hermana le queda pequeño, y quiere entrar en él.* She wants to l_____ some w_____ by next month.
29.	*Mi hermana vive en Nueva Zelanda, pero nos mantenemos en contacto muy frecuentemente gracias a redes de internet como Facebook, Skype o Twitter, que se han hecho muy populares.* These types of s_____ n_____ have grown incredibly popular in the last decade.
30.	*Me gusta saber para qué es la medicación que me mandan, por eso siempre le pregunto a mi médico.* I think they should explain the reasons for p_____ any m_____.
31.	*Mi mujer y yo hemos comprado una casa en ruinas y la hemos arreglado entera para hacerla habitable. Hemos tenido que hacer trabajos muy difíciles, pero ha merecido la pena.* It has been very h_____ w_____, but it was worth it.
32.	*Brian no recibió la beca de estudios este año porque el salario anual de sus padres era muy alto.* His parents´ a_____ i_____ was too large.
33.	*Ayer por la noche hubo una tormenta enorme, y los truenos no me dejaron dormir.* There was a big storm l_____ n_____.
34.	*Como tengo que hacer el trabajo de fin de máster durante el verano no tengo vacaciones, pero cuando puedo hago una excursión de un día con mis amigos a ciudades que están cerca.* Last week the d_____ t_____ was to London.

35.	*Las personas nos diferenciamos de los animales en que ellos carecen de racionalidad.* Consequently, we can say it is a unique characteristic of h_____ b_____.
36.	*Los votantes siempre se sienten atraídos por políticas de creación de empleo.* Therefore, politicians always promise to g_____ more j_____.
37.	*Cuando estaba en el colegio cada alumno tenía un casillero donde podía dejar sus objetos personales a salvo.* In high school we were careful to keep our p_____ b_____ safe.
38.	*El paisaje de Irlanda en verano ofrece tal variedad de colores y tan bonitos que te dejará sin palabras.* The w_____ r_____ of colours is beautiful.
39.	*El novio de mi amiga es muy alegre y siempre ve el vaso medio lleno.* He seems to have a p_____ a_____ towards life, and I envy him for that.
40.	*Cristina cree que es parte de la naturaleza de las personas el envidiar a la gente que tiene lo que tú deseas.* She thinks that envy is just h_____ n_____.
41.	*Trabajo muchas horas al día y no me queda tiempo para hacer lo que más me gusta y disfrutar de la vida. Quiero cambiar esto en el futuro.* I want to find some more f_____ t_____ next year.
42.	*La amiga de mi madre fue despedida de la empresa donde trabajaba y estuvo en el paro más de 6 meses.* Eventually, she f_____ another j_____, and it was much better than the previous one.
43.	*Aunque tengo una opinión muy clara en cuanto al aborto, es un tema muy controvertido y prefiero no hablar de ello.* Sometimes it is wise not to speak about c_____ i_____ like abortion.
44.	*Mis padres estuvieron juntos durante 25 años, pero hace ya 3 que se divorciaron.* It has been 3 years since my parents were a m_____ c_____.
45.	*Creo que el camarero no es de este país. Lo noté cuando habló con nosotros.* Yes, he spoke with a bit of a f_____ a_____.
46.	*Corea del Norte e Iraq dicen poseer armas de destrucción masiva, y así amenazan a*

	sus enemigos. Some countries are thought to possess n_____ w_____, and this is dangerous.
47.	*Para ser sincero, me atrae la idea de vivir en otro país durante algún tiempo.* I wouldn´t mind to l_____ a_____ for some time.
48.	*La fiesta de William fue perfecta hasta que a las 2 de la mañana vinieron los vecinos diciendo que dejásemos de hacer ruido porque su bebé no podía dormir.* William´s neighbours came and ask us to k_____ q_____.
49.	*Dicen que el asesino tiró la pistola al río e intentó quemar los cuerpos para deshacerse de las pruebas.* Obviously, he was trying to d_____ the e_____.
50.	*En España las rotondas se cogen en el sentido contrario a las agujas del reloj, pero no es así en todos los países.* In the UK, for example, they do it in a c_____ d_____.

Appendix 3 Language Background and Use Questionnaire

Para ayudarnos a entender, interpretar y clasificar mejor tus respuestas, ¿te importaría contarnos un poco sobre ti y tu experiencia en el aprendizaje de idiomas? Por favor, proporciona la siguiente información poniendo un tick (✓) en el recuadro o escribiendo tu respuesta en el hueco.

[In order to help us to better understand, interpret and classify your answers, would you mind telling us more about your personal and language learning background? Please provide the following information by ticking (✓) in the box or writing your response in the space.]

Gender: ☐ Male ☐ Female

Age: _____

How many years have you been studying English? _____

Which is your level of English? ☐ Beginner ☐ Intermediate ☐ Advanced

Are you studying English at the moment? ☐ Yes ☐ No

Have you spent a long period (3 months or more) in English-speaking countries? ☐ Yes ☐ No

How much time per week do you spend…:

- **reading books, magazines and newspapers in English, or visiting English language websites?** ☐ 0-1 hours ☐ 1-2 h. ☐ 2+ h.
- **watching films, videos or TV in English?** ☐ 0-1 hours ☐ 1-2 h. ☐ 2+ h.
- **listening to music in English?** ☐ 0-1 hours ☐ 1-2 h. ☐ 2+ h.
- **using English to keep in contact with people? (Facebook, MySpace, Twitter, Skype, email, SMS, etc.)**: ☐ 0-1 hours ☐ 1-2 h. ☐ 2+ h.

Finalmente, nos gustaría agradecerte mucho tu cooperación. Apreciamos muchísimo tu ayuda y contribución a este estudio. ¡Muchas gracias! ☺

Si estás interesado en recibir información sobre los resultados de este estudio, por favor, no dudes en contactarme por email en: _____@gmail.com

[Finally, we would like to thank you very much for your cooperation! We appreciate your help and contribution to this survey a lot. Thank you! ☺

If you are interested in receiving information about the results of this survey, please feel free to email me: _____@gmail.com]