The Acquisition of Phrasal Vocabulary by Non-Native Speakers of Spanish

A thesis submitted in partial fulfilment of the requirements for the

Degree of Master of Science in Psychology

in the

University of Canterbury

by

Victoria Escaip
2008

Research Supervisors:

Professor Kon Kuiper
Professor Garth Fletcher
# TABLE OF CONTENTS

List of Figures ........................................................................................................ 6

List of Tables ........................................................................................................... 6

Acknowledgements .................................................................................................. 7

ABSTRACT ............................................................................................................... 8

INTRODUCTION .................................................................................................... 10

1.1. In Search of a Language Model ................................................................. 11

1.1.1. The creative nature of language ......................................................... 11

1.1.2. The open choice principle versus the idiom principle ............ 13

1.1.3. A dual model of language ability ...................................................... 14

1.1.4. An integrated model of language ....................................................... 16

1.2. What is Phrasal Vocabulary? .................................................................... 18

1.2.1. Identifying phrasal vocabulary ......................................................... 18

1.2.2. Non-compositionality of formulaic sequences ......................... 19

1.2.3. Fixedness of formulaic sequences .................................................. 20

1.2.4. Functions of formulaic sequences ................................................... 21

1.3. Formulaic Language Acquisition ......................................................... 23

1.3.1. The acquisition of phrasal vocabulary by native speakers ....... 23

1.3.2. The acquisition of phrasal vocabulary by non-native speakers .... 24

1.3.3. Awareness and the acquisition of formulaic sequences by non-native speakers ......................................................... 26
1.3.4. Further obstacles in the acquisition of formulaic language
by non-native speakers...................................................... 28
1.4. Cultural Integration......................................................... 29
  1.4.1. Relationship between culture and formulaic language......... 29
  1.4.2. Research on cultural integration and formulaic language
          acquisition by second language learners......................... 31
1.5. Corpora Exploration of Formulaic Sequences......................... 33
  1.5.1. Advantages of corpus research.................................. 33
  1.5.2. Problems of corpus analysis..................................... 35
1.6. Methodological Issues................................................... 37
  1.6.1. Superlemma Theory............................................... 37
  1.6.2. Cloze test methodology......................................... 39
  1.6.3. Superlemma Theory and cloze testing.......................... 42
1.7. Research Rationale...................................................... 43
  1.7.1. Antecedents: Acquiring phrasal vocabulary.................... 43
  1.7.2. The present study: The acquisition of phrasal vocabulary
          by non-native speakers of Spanish................................. 44

METHOD................................................................. 46

  2.1. Participants.......................................................... 46
  2.2. Measures.............................................................. 47
    2.2.1. Cultural questionnaire.......................................... 47
    2.2.2. Cloze test...................................................... 48
  2.3. Procedure............................................................. 53
RESULTS

3.1. Differences between Scores of Native Speakers and Non-Native Speakers of Spanish in the Cloze Test

3.2. Is the Frequency of the Head-Verb of a Formulaic Expression Linked with the Knowledge of Related Expressions?

3.3. Links between Age and the Mastery of Formulaic Expressions in Spanish for Native Speakers

3.4. Links between Time, Age and Culture, and the Mastery of Formulaic Expressions in Spanish for Non-Native Speakers

3.5. Is Cultural Integration a Significant Predictor for the Mastery of Formulaic Expressions?

DISCUSSION

4.1. Summary of the Results

4.2. The Research Hypotheses

4.2.1. There are significant differences between the degree of acquisition of formulaic sequences in native and non-native speakers of Spanish

4.2.2. The frequency of usage of the head-verbs contained in verb plus complement formulaic sequences is tightly linked to the acquisition of such sequences

4.2.3. Phrasal vocabulary of native speakers is age graded in that much of it is acquired in late adolescence and adulthood

4.2.4. Cultural integration is closely linked to the acquisition of formulaic language
4.3. Cross-Cultural Implications......................................................... 80
4.4. Applications.................................................................................. 82
    4.4.1. Teaching formulaic language............................................... 82
    4.4.2. The importance of word frequency in vocabulary teaching..... 85
4.5. Strengths and Limitations................................................................. 87

CONCLUSION.......................................................................................... 89
REFERENCES........................................................................................... 91
APPENDICES........................................................................................... 98
List of Figures

Figure 1: Cutting & Block (1997) Hybrid Model in Sprenger, Levelt & Kempen(2006)…………………………………………………………38

Figure 2: Sprenger, Levelt & Kempen (2006) Superlemma Model…………. 39

Figure 3: Means of the Number of Correct Answers for the Four Categories of Formulaic Expressions According to Their Head-Verb Frequency…………………………………………………………58

List of Tables

Table 1: Means and standard deviations for the scores in the cloze test by both groups of participants……………………………………56

Table 2: Means and standard deviations of the number of correct answers in the cloze test for both groups of participants……………………………………………………………………………………………………58

Table 3: Correlation coefficients for native speakers of Spanish between the major variables……………………………………………………60

Table 4: Correlation coefficients for non-native speakers of Spanish between the major variables…………………………………………………61
Acknowledgments

I would like to acknowledge the invaluable contribution of Professor Kon Kuiper for his guidance, advice and encouragement during the completion of this thesis. I would also like to thank Professor Garth Fletcher who provided important additional comment and statistical advice.

I also thank my family and close friends, who have always been a source of love and inspiration.

Finally, I would like to express my deep gratitude to my husband Alfredo Escaip for all his devotion, patience and loving support.
ABSTRACT

The term ‘phrasal vocabulary’ refers to multi-word expressions, that is, idioms, templates or “strings of words, which appear to be processed without recourse to their lowest level of composition” (Wray, 2002, p.4). Formulaicity constitutes an essential feature of language production and comprehension, and phrasal vocabulary plays a central role in everyday language usage.

This research study replicates the experimental design carried out in the study Acquiring phrasal vocabulary by Kuiper, Columbus, & Schmitt (to appear), which used a cloze procedure to test three main hypotheses: a) There are significant differences between the degree of acquisition of formulaic language items by native and non-native speakers of English; b) The frequency of usage of the head-verbs contained in verb plus complement formulaic sequences is positively correlated with the acquisition of such sequences; and, c) Phrasal vocabulary is age graded. In the present study the target language is Spanish instead of English.

In addition, available evidence suggests that cultural integration seems to be linked to the acquisition of formulaic language. Thus, a questionnaire intended to measure the participants’ cultural integration level to the target language community was developed.

The results of this study supported the predictions that the amount of formulaic language acquired by native speakers is positively correlated with age, and that non-native speakers’ phrasal vocabulary is significantly less extensive than that of native speakers. Most importantly, the results also showed a significant effect of verb frequency on the participants’ acquaintance with the formulaic sequences tested.
However, the prediction that cultural integration would be positively correlated with the number of correct answers in the cloze test for both groups was not supported.

Extending to the Spanish language the results reported by Kuiper, Columbus, & Schmitt supports the argument that the processes of acquisition of formulaic language across diverse linguistics systems function in a very similar way (Corpas Pastor, 2003). A better comprehension of the mechanisms by which speakers acquire formulaic language may significantly contribute to the development of an appropriate methodology to teach phrasal vocabulary to second language learners.
INTRODUCTION

“You shall know a word by the company it keeps.”

(Firth, 1968, p.179)

It is clear that a large part of human language production and comprehension comprises ‘prefabricated phrases’ that often have a unique or distinctive meaning, which cannot be explained through the analysis of the different lexical items from which they are composed. When engaged in the production of speech, this capacity gives people rapid access to ‘chunks’ of information that have been ‘pre-packaged’ into single units, thus increasing the proficiency of verbal communication. These units constitute what is called phrasal vocabulary or formulaic language, and represent an important trait of language production from the perspective of both the speaker and the listener (Kuiper, Columbus, & Schmitt, to appear; Schmitt & Carter, 2004; Wray, 2002; Wray & Perkins, 2000).

Formulaicity is ubiquitous in human languages. For instance, studies have shown the likelihood of establishing the same general criteria when exploring formulaicity in the English and Spanish languages, and possibly in other Romance and Germanic languages (Corpas Pastor, 1996, 2003). This is the reason why English and Spanish phraseology will not be treated as separate fields of study in this research, and the theoretical and empirical considerations on the language processes of acquisition, storage and comprehension will be taken to apply to both languages.
1.1. In Search of a Language Model

1.1.1. The creative nature of language

It is undeniable that human language is the result of intricate processes that operate in the human’s mind. Jackendoff (1994) claims that the ability to speak and understand language requires the capacity to store in the mental lexicon not only words, but also patterns of words, which subsequently explain patterns of patterns in our language. These patterns represent the rules of language that constitute the ‘mental grammar’, which allows us to make sentences from single words. The child acquires this system of rules when learning how to speak, but this grammar is not available to the consciousness of the speaker, and only its output is (Jackendoff, 2007). The mental grammar is constituted not only by an ‘Innate Part’ or ‘Universal Grammar’, but also by a ‘Learned Part’. Consequently, language ability is the result of a combination of genetic and environmental factors, where the learner is the active agent of learning, and education (among a number of other influential factors of the environment) serves as a commitment and resourcefulness magnifier (Jackendoff, 1994).

Chomsky (1965) emphasizes the creative nature of language which enables humans to formulate brand new sentences never heard or spoken before. He proposes a single processing system that is the product of inherent mind/brain structures and bestows on people a great analytical capability for the understanding, acquisition and generation of language. The genetic hypothesis proposes that “the ability to learn language is rooted in our biology, a genetic characteristic of the human species, just like an opposable thumb and a pelvis adapted for upright stance” (Jackendoff, 1994, p.30). Linguists have proposed over time that this system of rules allows people to
construct countless novel sentences from a limited grammar (Van Lancker-Sidtis & Rallon, 2004). In line with this, Pinker argues that “virtually every sentence that a person utters or understands is a brand-new combination of words, appearing for the first time in the history of the universe” (1995, p.22). Thus, from this point of view, human language is characterised by its novelty, and this essential property provides people with the resources for expressing their thoughts in an indefinite variety of circumstances (Chomsky, 1965).

However, this inherent analytical faculty for processing language that enables humans to comprehend and produce utterances that they have never come upon before, has been largely overestimated, whereas humans’ great memory capacity and their ability to produce language using ‘prefabricated units from memory’ have been undervalued (Lamb, 1998). Even though it is irrefutable that humans possess this great cognitive capacity for processing language, a number of researchers have proposed that a great deal of language comprehension and production is not only accomplished analytically. It has been argued that humans also manage to create and understand language using prefabricated chunks of speech that are retrieved from memory as a whole unit (Jackendoff, 1995; Kuiper, Columbus, & Schmitt, to appear; Lamb, 1998; Pawley & Syder, 1983; Sinclair, 1987, 1991; Sprenger, Levelt, & Kempen, 2006; Van Lancker-Sidtis & Rallon, 2004; Wray, 2002; Wray & Perkins, 2000).

Several models of language acquisition have been proposed over the years. The Chomskian tradition has prevailed in the formulation of many of them and, although many researchers believe that prefabricated expressions play a definite role in language learning, some consider the importance of these expressions as peripheral to the processes of language (Krashen & Scarcella, 1978). However, formulaicity has
been strongly recognised as an important attribute of language in modern linguistics and psycholinguistics, and numerous investigations on the role of formulaic language in the mechanisms of language have been carried out in the last few decades. Some of the relevant models of language that account for this essential feature of language are presented below.

1.1.2. The open choice principle versus the idiom principle

Sinclair (1987; Sinclair, 1991) explores the means by which people handle linguistic material, and proposes the existence of two models of language comprehension: the open choice principle versus the idiom principle. The open choice principle entails that understanding language is the result of a number of a large range of choices only limited by grammar, whereas the idiom principle offers the language users sets of linguistic choices, i.e., “a large number of semi-preconstructed phrases that constitute single choices, even though they appear to be analysable into sections” (Sinclair, 1987, p.320). Thus, the open choice principle operates through the selection of single words, while the idiom principle leads to the selection of blocks made up of two or more words that have previously occurred together in a regular way (Sinclair, 1991).

An essential attribute of the idiom principle is that it reflects the natural human tendency to economy of effort (Sinclair, 1987). The use of lexical phrases provides the speaker with fluency since the speed for processing them as whole units (as if they were single lexical items) increases the efficiency and reduces the time of retrieval of language items from memory (Nattinger & DeCarrico, 1992). Basically, the use of formulaic sequences eases the mind’s processing load as these multi-word expressions
are processed more easily and faster than the same series of words generated under the novel language scheme (Kuiper, Columbus, & Schmitt, to appear).

Corpas Pastor (1996) emphasises the importance of formulaic sequences within the lexical component of languages and the linguistic production of native speakers, because this efficiency provides the latter with sufficient time for planning longer units of speech and overseeing the social aspects of communication.

An interesting question is that of Wray’s (2002) about what could be the processing advantage of storing holistically some long formulaic sequences, when the same message they convey could be expressed in a couple of words. She speculates that long expressions are used to buy time for thinking, and also to facilitate an even rhythm in the conversation. Thus, economy of effort is not “simply about taking short cuts [but also] regulating production” (p. 75).

Sinclair (1987) highlights the incompatibility of these two models, the open choice principle and the idiom principle, and refers to them as ‘diametrically opposed’. He claims that the importance of the principle of idiom has been largely neglected, but that it is precisely this principle which fundamentally explains the processes involved in the acquisition, comprehension and generation of language. In his view, the idiom principle dominates, while the open choice principle is a secondary model.

1.1.3. A dual model of language ability

Phrasal vocabulary is ubiquitous and it is hard to have an approximate estimation of the extent of the formulaic language stored in the mental lexicon (Kuiper, Columbus, & Schmitt, to appear). Pawley & Syder (1983) claim that phrasal vocabulary prevails in the speakers’ output, while novel sentences constitute just the
minority. Jackendoff (1995) suggests that the number of fixed expressions stored in
the mental lexicon is vast. Some estimates suggest that it is likely to be ten times as
large as the single word lexicon: “In the lexicon, phrasemes are more numerous than
words by a ratio of at least 10 to 1” (Mel'čuk, 1995, p.169). Many linguistics scholars
agree that prefabricated phrases amount to tens of thousands (Van Lancker-Sidtis &
Rallon, 2004). Thus, the available evidence on the prevalence of prefabricated strings
of words in language users makes obvious the implausibility of describing and
explaining the processes of language acquisition, production and comprehension, by a
single-system model (Wray, 2002).

Furthermore, available evidence of neurological studies establishes the
interplay of automatic and novel processes involved in many types of human
behaviour (Van Lancker-Sidtis & Rallon, 2004). There is evidence from
psycholinguistics studies on the language of normal and brain-damaged individuals
that reveals that novel sentences and formulaic sequences are processed by different
neurological structures, and represented in different cerebral hemispheres (Van
linguistic skills seem to reside in the left hemisphere, while phrasal vocabulary and
single word lexicon with similar properties of use are processed in the right
hemisphere of the human brain (Kuiper, 2006).

Van Lancker & Rallon’s (2004) compared the incidence of formulaic
expressions versus novel expressions in the text of the screenplay Some Like It Hot.
They found that formulaic expressions, categorised either as speech formula, idiom or
proverb, made up nearly a quarter of the sentences in the text. Then, in a verification
survey to obtain an indication of the proportion of persons agreeing with their
identification of fixed expressions, the participants performed significantly higher on
formulaic expressions than on novel phrases, which reflected the general knowledge of the former. Their findings revealed the prevalence of formulaic expressions in everyday speech of adult speakers, and supported the existence of a dual model of language ability that entails both holistic and analytical processes, i.e., a model that alternates ‘fixity’ and ‘creativity’ when processing language (Tannen, 1989).

This dual model of language ability is characterised by the interplay of the holistic process that allows humans to cope with language by bringing about a vast number of prefabricated sentences, and the analytical process by which speakers handle language using their syntactic skills (Van Lancker-Sidtis & Rallon, 2004).

1.1.4. An integrated model of language

Wray and Perkins’ (2000) integrated model of language involves the two strategies for processing language mentioned above: analytic processing to produce and decode novel language that implies the interaction of single lexical items with the rules of grammar, and holistic processing that depends on pre-assembled strings of words stored in the mental lexicon, i.e., formulaic sequences. Their model assigns the central role in communicative language processing to the holistic system, but it does not deny at any time the importance of the analytical or creative system. The establishment of a suitable balance between the two is proposed as ‘the best deal’, since seeing either of those two ways of dealing with language as exclusive would be restrictive. Thus, the analytical or creative scheme enables language users to generate or decode the unforeseen, whereas the holistic system provides speakers processing effort economy by making ready-made utterances available to them in contexts which are predictable (Wray, 2002).
According to Wray and Perkins (2000), the selection of a particular strategy for processing language in adult speakers will be determined by the priorities of social interaction and the limitations of the language user’s memory capacity. The holistic strategy will represent the first choice as a result of the natural tendency of humans to economise effort, or due to the time constraints during the conversation (Sinclair, 1987). It also constitutes the best alternative when the speaker requires concentrating on a particular subject that is different from the conversation’s matter. Kuiper (1996) declares that trading off processing effort against creativity has many advantages in a range of situations. “Formulae make the business of speaking (and that of hearing) easier … When a speaker uses a formula he or she needs only to retrieve it from the [internal] dictionary instead of building it up from its constituent parts. In other words, such expressions likely exist as whole or part utterances within the speaker’s dictionary and need not be built up from scratch on every new occasion” (p. 3).

Therefore, formulaic structures are primarily selected for language output, but the analytical system is always available to assist language users to solve any production and/or comprehension language problems that arise from the unexpected. An integrated model of language explains the “moment-by-moment strategy choices of the individual” by successfully decoding novel input, processing language efficiently, and ensuring the understanding of the message (2000, p.18).
1.2. What is Phrasal Vocabulary?

“Where is the boundary between a relatively independent item and one with such a strongly determining environment that we are tempted to extend the item boundary and recognise a phrase?" (Sinclair, 1996, p.82)

1.2.1. Identifying phrasal vocabulary

Phrasal vocabulary is comprised of all the multiword expressions that are stored holistically in the mental lexicon of the language users, that is, “multi-lexemic expressions which are perceived, memorised and retrieved as single units and are thus processed as indivisible wholes” (Munat, 2002, p.145). These sequences, or strings of words, constitute what is known in more general terms as formulaic language (Corpas Pastor, 1996, 2003; Kuiper, 1996, 2006; Kuiper, Columbus, & Schmitt, to appear; Nattinger & DeCarrico, 1992; Ruiz Gurillo, 2001; Schmitt & Carter, 2004; Van Lancker-Sidtis & Rallon, 2004; Wray, 2002; Wray & Perkins, 2000).

There is not unanimous agreement among researchers on what is considered as formulaic language, and consequently there is not a single satisfactory definition for this phenomenon. However, formulaicity may be defined as a phenomenon “manifested in strings of linguistic items where the relation of each item to the rest is relatively fixed, and where the substitutability of one item by another of the same category is relatively constrained” (Wray & Perkins, 2000, p.1).

Formulaic expressions have been defined in a number of ways. According to Wray (2002; Wray & Perkins, 2000) more than 50 terms have been used to refer to prefabricated phrases or strings of words that are considered to be within this
category. Among them are *amalgams, chunks, clichés, collocations, complex lexemes, composites, conventionalised forms, fixed expressions, formulas/formulae, frozen phrases, idioms, lexical phrases, multiword items/unit, prefabricated routines and patterns, ready-made expressions/utterances*, etc. (Wray, 2002, p.9).

For the purpose of this study, Wray’s (2002) definition of formulaic sequences will apply as it includes two principles that are central to the body of this research: their multi-word character, and holistic nature (Schmitt & Carter, 2004). Wray defines the formulaic sequence as “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (p.9).

Formulaic language is extremely versatile and proof of this is the diversity of terms and definitions that have been used to define it. Formulaic expressions are difficult to classify as they often fit in more than one category (Van Lancker-Sidtis & Rallon, 2004). Since formulaicity is multifaceted and cannot be described on the basis of a single criterion, researchers have explored several aspects of formulaic language in order to get a more comprehensive understanding of this phenomenon. Some of its main features are examined next.

1.2.2. Non-compositionality of formulaic sequences

It has been suggested that some phrases that are originally constructed using the rules of grammar, can develop into formulaic sequences as a consequence of their frequent use in the speech (Corpas Pastor, 1996; Peters, 1983; Ruiz Gurillo, 2001) Thus, in spite of their transparency, they are not anymore a product of analytical
processes since they have become preconstructed lexical chunks that are accessible to
the speaker, and the listener, through the holistic processes of language.

One of the areas of idiosyncrasy common to many formulaic expressions is
their non-compositionality, that is, their whole meaning cannot be deduced from the
meaning of each of their separate lexical constituents. This non-compositionality, or
opacity, which originates for several reasons through the evolution of the language
practices in a particular community, is considered as one of the main characteristics of
formulaic sequences (Corpas Pastor, 1996; Ruiz Gurillo, 2001; Wray, 2002). Opaque
formulaic expressions may be difficult for a listener to understand, if not impossible,
when heard for the first time without previous pragmatic or contextual knowledge
(Wray, 2002; Wray & Perkins, 2000). Ruiz Gurillo (2001) states that literal
translation of opaque formulaic sequences is virtually impossible. This opacity
prevents second language speakers from analysing them and, consequently,
comprehending them. She says that the understanding of an opaque formulaic
expression is “like a buried treasure in a desert island” (p. 21). There are no easy
access roads to it.

It has been proposed that when describing formulaic expressions, it may be
helpful to conceptualise them on a continuum based on their degree of their
opacity/transparency, “from fully bound to fully free” (Wray, 2002, p.34).

1.2.3. Fixedness of formulaic sequences

Even though some researchers suggest that fixedness is an essential
characteristic of formulaic sequences, it is clear that only a small group of these
expressions is totally fixed, and the rest of them contain slots that allow the legitimate
insertion of a variety of elements (Wray, 2002). In addition, speakers creatively
modify formulaic sequences all the time, but the high level of institutionalisation of these expressions enables the language users to still recognise them (Corpas Pastor, 1996).

According to Corpas Pastor (1996), the degree of modification of formulaic expressions is positively correlated to their degree of fixedness. That is, the more institutionalised a sequence is, the more possible to be modified without changing its original meaning.

Fixedness and non-compositionality (opacity) are the two most generally accepted criteria for the identification of formulaic sequences (Read & Nation, 2004; Ruiz Gurillo, 1998). Ruiz Gurillo (1995; Ruiz Gurillo, 1998) proposes that the process of ‘fraseologización’¹, through which a formulaic sequence is generated, is determined by a certain degree of fixedness, and, on many occasions, by a partial or total non-compositionality. “The fact that these criteria turn out to be continua contributes to the difficulty in drawing the line between formulaic and non-formulaic expressions” (Read & Nation, 2004, p.33).

1.2.4. Functions of formulaic sequences

Formulaic sequences are resources available to language users who employ them depending on a number of factors such as maturation, language comprehension and social interaction needs. Formulaic language benefits the speaker by assisting the speaker’s production and the hearer’s comprehension (Wray, 2000, 2002; Wray & Perkins, 2000). Formulaicity not only renders the language production process efficient, but also provides socially appropriate frames for communicating, since countless formulaic sequences are closely related to particular language functions in

¹ This Spanish term that could be translated into English as ‘phraseologisation’, is used to refer to the processes involved in the generation of a phrase which is considered as a formulaic sequence.
social interaction (Schmitt & Carter, 2004; Wray, 2002; Wray & Perkins, 2000).
Although these functions can also be achieved using novel language, the association
between formulaic sequences and language functions in social interaction is evident in
the following ways:

- Changing one’s physical and perceptual environment to satisfy physical,
  emotional and cognitive needs using: commands (Keep off the grass); requests
  (Could you repeat that please?); politeness markers (I wonder if you mind...);
  and, bargaining (I’ll give you _____ for it).
- Asserting separate identity to be taken seriously employing: story telling
  (You’re never going to believe this, but...); and, turn claimers and holders
  (Yes, but the thing is...). Also to be separated from the crowd using personal
  turns of phrase (I wanna tell you a story).
- Asserting group identity to acquire overall membership utilizing: ‘In’ phrases
  (Praise the Lord!); group chants (We are the champions); institutionalised
  forms of words (Happy birthday: clearly beloved, we are gathered here
today...); and, ritual (Our Father, which art in Heaven...). And to achieve as
  well a place in hierarchy (affirming and adjusting) using: threats (I wouldn’t
do that if I were you); quotation (“I wouldn’t want to belong to any club that
would have me as a member” (Groucho Marx)); forms of address (Your
Highness); and, hedges (Well I’m not sure) (Wray, 2002, p.89).

From an evolutionary point of view, Wray & Perkins (2000) propose that the fact
that these functions also feature in the ‘holistic noise-gesture communication of
primates’, points to a parallelism that may indicate that human language evolved from
a holistic protolanguage where utterances produced in the absence of a system of rules were associated with particular meanings (p. 14).

1.3. Formulaic Language Acquisition

1.3.1. The acquisition of phrasal vocabulary by native speakers

Kuiper, Columbus, & Schmitt (to appear) claim that the study of phrasal vocabulary has concentrated on its detection and description, rather than into the mechanisms of its acquisition. Research has focused on longitudinal accounts of language development, rather than in the investigation of the processes that explain how language users acquire and store formulaic language.

It has been proposed that children may build up generative grammar by segmenting the formulaic expressions already acquired into smaller components (Kuiper, Columbus, & Schmitt, to appear; Peters, 1983; Wray, 2002). However, it has also been argued that children segment certain sequences only when this segmentation is potentially useful for social communication purposes. For the most part, children will tend to process language in a holistic way and will produce a number of formulaic sequences that will be retained into adulthood.

Wray & Perkins suggest a four phase developmental model where a significant amount of formulaic language is acquired in early childhood, then oddly vanishes for a few years, and reappears during adolescence, increasing consistently towards adulthood (Wray, 2002; Wray & Perkins, 2000). This model can be summarised as follows (Schmitt & Carter, 2004):
In phase 1, which starts at birth, the child basically uses holistically processed language by identifying, selecting, and storing formulaic sequences that are socially meaningful within the linguistic environment. These sequences will subsequently allow the activation of the analytical processes of language in phase 2. During phase 2, which starts between 20 and 30 months of age and lasts until the age of 8 years, the child will show a preference for using analytical mechanisms to process language. This is when grammatical awareness starts, and throughout this period the child will analyse and re-combine the constituents of the expressions previously acquired using the rules of grammar. Analytic language will then prevail, but the amount of formulaic language will still show an increase since the child’s language is developing on the whole. It is in Phase 3, which goes until around the age of 18, when the analytical system has been firmly developed, and formulaic language recovers its predominance over analytical language. By phase 4, a balance between both systems of language processing, analytic and holistic, has been achieved, and adult patterns of formulaicity are established.

However, as Kuiper, Columbus, & Schmitt (to appear) point out, there is not yet an explanation that describes the means by which people acquire phrasal vocabulary in the first instance.

1.3.2. The acquisition of phrasal vocabulary by non-native speakers

As Wray (2002) puts it: “To know a language you must know not only its individual words, but also how they fit together” (p. 143). Proper knowledge and use of formulaic language is vital to sound like a native speaker (Pawley & Syder, 1983; Van Lancker-Sidtis & Rallon, 2004; Wray, 2000; Wray & Perkins, 2000).
On the whole, evidence reveals that second language learners have great difficulties acquiring and producing native-like formulaic language. In contrast to native speakers who process language largely by means of the holistic system, non-native speakers may seem to concentrate more on single words than on strings of words (Kuiper, Columbus, & Schmitt, to appear). It seems that only a few non-native speakers fully acquire the native collection of formulaic sequences (Pawley & Syder, 1983).

The acquisition of formulaic sequences by non-native speakers may not necessarily follow the same patterns that occur in the process of acquiring phrasal vocabulary by native speakers. It has been proposed that in the early stages of second language learning many useful formulaic sequences develop quickly into the vocabulary of either classroom-taught or naturalistic adult learners. The main purpose would be to achieve success in communicating with others within their linguistic environment (Nattinger & DeCarrico, 1992; Wray, 2002; Wray & Perkins, 2000).

However, in later stages of learning, the acquisition of formulaic sequences, and consequently their use, falls behind expectations. It appears that second language learners rely a great deal on the creative language processes, and tend to segment formulaic sequences through the grammatical analysis of their components by applying the grammar rules (Rod Ellis, 2005; Nattinger & DeCarrico, 1992; Wray, 2002). This analytical process leads second language learners to identify within the holistically learned sequences single lexical items that are learned independently, and then perform, partially or fully, an inevitable segmentation (Wray, 2002). Second language learners “tend to over-generate, producing grammatical utterances that are simply not idiomatic” (Wray & Perkins, 2000, p.23). Therefore, proficient learners
often produce flawless sentences constructed using the rules of grammar that native speakers would never use (Wray, 2002).

On the other hand, Nattinger and DeCarrico (1992) maintain that this segmentation process, which is executed on analogy with similar sequences, is a valuable tool for language learners who “break these chunks down into sentence frames that contain slots for various fillers” (p. 115), expanding significantly their ability to communicate. These authors consider that lexical phrases are an ideal unit for language teaching as they are central to the processes of the acquisition, comprehension and generation of language. Ellis (2005) shares this view and claims that the acquisition of formulaic sequences constitutes a foundation for the “development of a rule-based competence” (p. 211). However, Wray (2000) argues that there is an incongruity within this essentially analytic approach to teaching formulaic sequences, since the ‘very nature’ of these expressions is fundamentally holistic: they are learned as whole units and are not analysed.

Thus, when comparing the patterns of formulaic language acquisition by non-native speakers with those of native speakers, it appears that second language learners linger in phase 2 or 3, and they do not go further to phase 4 where the balance between creative and formulaic language settles (Wray & Perkins, 2000).

1.3.3. Awareness and the acquisition of formulaic sequences by non-native speakers

Bishop (2004) believes that a reason of the difficulty of adult second language learners for acquiring formulaic language may be the lack of awareness of this phenomenon. He claims that the non-detection of the holistic form of idiomatic expressions hinders the capacity of the speakers in learning and storing these sequences as single units. This author emphasises that in order to learn a formulaic
expression one must recognise it first. However, while unknown single words are perceived more easily because they are ‘clearly delineated’, native speakers have problems in recognising formulaic expressions due to their inconsistent shape. In most cases formulaic sequences are non-compositional and it is difficult for non-native speakers to recognise them as whole linguistics units (p. 229). Bishop investigated the effects of typographic salience of formulaic sequences in a text on the readers’ eagerness to find out their meaning, and the contribution of this glossing to these expressions being learned, obtaining in both cases only moderate positive correlations.

Jones & Haywood (2004) also explored the awareness of non-native speakers in a study with pedagogical purposes. They highlighted a range of formulaic sequences during a ten-week course for EAP (English for Academic Purposes) students, through repeated exposure and discussion. The results revealed an increased awareness of the unitary nature of formulaic sequences, but this awareness only produced a minor enhancement in the students’ production of this kind of expressions, and an even smaller improvement in the production of phrasal vocabulary in their essays.

Another study related to the subject of awareness is Van Lancker’s (2003) research that examined the phonological form of formulaic sequences in spoken language, and investigated the auditory recognition of idioms by highly proficient second language speakers. In a preceding study (Van Lancker-Sidtis & Canter, 1981), it was found that native speakers of English were able to discriminate between idiomatic and literal sentences using prosodic cues. The results of Van Lancker subsequent research indicated that fluent non-native speakers scored significantly lower than native speakers in distinguishing sentences with literal meanings from
sentences with idiomatic meanings in a listening task. She proposes that the acquisition of formulaic expressions and their prosodic characteristics may be acquired together by native speakers, and that these prosodic cues are language specific (Van Lancker-Sidtis, 2003).

1.3.4. Further obstacles in the acquisition of formulaic language by non-native speakers

It has been proposed that the speech addressed to second language speakers does not contain much phrasal vocabulary and, therefore, learners simply do not have the required exposure to these sequences. Thus, when in the process of looking for an appropriate expression, adult second language speakers will have to ‘make their best guess’ (Wray & Perkins, 2000).

However, studies of naturalistic learners reveal that in many cases they will use their non-native status as a means to withdraw themselves from the communicative situation if they are not confident enough to handle it (R. Ellis, 1994). Second language speakers may also downsize their social interactional needs to make them fit with their actual knowledge of formulaic sequences (Wray & Perkins, 2000).

Evidence suggests that learners tend to avoid the use of these sequences, and they are inclined to include familiar expressions (made up analytically from single words) in their speech because they feel more confident with their use (Schmitt & Carter, 2004). In countless cases, besides their inadequate knowledge of language, non-native speakers often ignore the appropriate cultural ways required for social interaction. Thus, in order to fulfil their socio-interactional needs, they bring together a collection of formulaic sequences that alternate appropriate with inappropriate expressions, and interlanguage ones (Wray & Perkins, 2000). Second language
speakers will then use their native language if they find out it works, and will also
tend to use fused expressions that, although not being native-like, serve them to
achieve their communication purposes (Wray, 2002).

It has been observed that many of these unidiomatic expressions become
fossilised and are incorporated into the lexical repertoire of the speaker, who will use
them frequently, reducing the possibilities of using other more suitable alternatives. In
addition, it appears that frequency of use facilitates these unidiomatic sequences
developing into linguistic units that are stored holistically in the mental lexicon of the
second language speaker. Thus, although it might seem paradoxical, the patterns of
language acquisition by non-native speakers “may be more supportive to the model

1.4. Cultural Integration

1.4.1. Relationship between culture and formulaic language

“The culturally sanctioned forms of words symbolise the identity of the
society, and mastery of them is a mark of status and trust” (Wray, 2002, p.76). It is
undeniable that important cultural values are embedded in the language of any
linguistic community. The sociocultural aspects of the society determine to a great
extent the construction and the form of the speech used by the language users (Corpas
Pastor, 1996).

According to Coulmas (1979), the routine formulas “are [prefabricated]
expressions whose occurrence is closely bound to specific social situations and which
are, on the basis of an evaluation of such situations, highly predictable in a
communicative course of events” (p. 240). This author argues that the situational frames for these expressions are cognitive schemas, which include the necessary information that triggers their appropriate use in a particular situation. These conceptual structures represent the prototypical and conventional perception of the members of a linguistic community, and are a true reflection of their culture.

In the same line, Garcia-Page (1995) argues that phraseology is deeply rooted in the history of any linguistic community, tightly linked to its idiosyncratic values: many formulaic expressions represent either contemporary linguistic stereotypes, or ‘archaeological’ vestiges of the language (p. 155).

Corpas Pastor (1996) argues that it is precisely due to this characteristic that many formulaic expressions do not have equivalents in other languages, since the situations that are sanctioned by them are culture specific. An example that she offers is the Arabic expression Na `īman (God blesses you) which is addressed to somebody who has just taken a bath or cut their hair, while this situation does not need any particular comment in other languages such as English, Spanish, French or German (p. 176).

Formulaic language is pivotal in a community’s identity. The speakers aspiration to ‘sound like others’ makes it likely that speech communities will possess their own collection of particular formulaic sequences, since language users will store them holistically after repeatedly hearing them in the speech of others (Wray, 2002). The use of formulaic expressions constitute, thus, some kind of social grammar or linguistic etiquette that emerges from the aspiration of speakers to acquire a fluent, efficient and coherent discourse within a particular language community (Corpas Pastor, 1996).
“Language learning is culture learning … Native-like competence in a language is only possible if it is accompanied by the acquisition of an appropriate native culture” (Kuiper & Tan Gek Lin, 1989, p.304). Therefore, cultural integration should be expected to result in non-native speakers’ desire to sound like the native speakers of the target language community, with the subsequent enhanced learning of native-like expressions, i.e., formulaic language.

1.4.2. Research on cultural integration and formulaic language acquisition by second language learners

Interaction needs and the use of formulaic expressions seem to be correlated (Schmitt & Carter, 2004). On the basis of the significant effects of culture on formulaic language use, some studies have been carried out to investigate the relationship between cultural integration and formulaic language, and some evidence shows that there may be a significant relationship between the level of integration into the second language cultural environment, and the acquisition of formulaic expressions by second language learners.

Wong-Fillmore’s (1976) findings, cited in Schmitt & Carter (2004), revealed that among eight of the strategies used by five Mexican children in order to integrate to the target culture, three involved the learning and use of formulaic expressions.

Dörnyei, Durow, and Zahran (2004) believe that the acquisition of formulaic sequences “is a socially-loaded process” (p. 87), and second language learners must incorporate cultural elements of the target language community into their language learning process. In their opinion, many second language learners fail when learning a language because they consider this to be a separate process from learning the second language culture. The results of their exploratory study on the links between language
acquisition and cultural integration did not reveal significant correlations between individual differences and the extent of formulaic language acquisition, but suggested that gains in phrasal vocabulary correlate with language aptitude, motivation and sociocultural adaptation. For success in acquiring formulaic sequences, only high levels of the first two factors can compensate for the absence of acculturation, whereas high levels of the latter can countermand low levels of language aptitude and motivation.

Adolphs and Durow (2004) longitudinal study’s results on the impact of the three factors mentioned above suggest that improvement in the use of common formulaic sequences over time was greater for a student who was highly socially integrated compared to that of a student with cultural integration difficulties.

However, an exploratory study on the teaching of formulaic sequences to a group of non native EAP (English for Academic Purposes) students did not find a significant correlation to language aptitude, language motivation, and language attitudes (Schmitt, Dörnyei, Adolphs, & Durow, 2004).

Additional contradictory evidence on the links of cultural integration and the use of formulaic expressions is found in the results of Siyanova and Schmitt (2007), cited in Kuiper, Columbus, & Schmitt (to appear), which showed that the amount of exposure to the target language speaking environment did not increase the probability of second language speakers of using multi-word verbs. However, as Kuiper, Columbus, & Schmitt point out, it may not be about the amount of exposure what is significant, but the quality of exposure that second language learners experience when socially integrated into the native speaking community.
1.5. Corpora Exploration of Formulaic Sequences

1.5.1. Advantages of corpus research

The examination of large corpora is an empirical approach to linguistic research that has constituted an important source of information for the study of formulaic language (Read & Nation, 2004; Schmitt, Grandage, & Adolphs, 2004; Tamburini, De Santis, & Zamuner, 2002; Wray, 2002). Corpus analysis allows researchers to produce frequency counts as well as other quantitative measures of word strings. It has also provided important information for the execution of qualitative analyses of formulaic sequences (Read & Nation, 2004; Schmitt, Grandage, & Adolphs, 2004). In addition, corpus research has been a valuable tool in lexicography and has permitted the compilation of major language dictionaries (Corpas Pastor, 2003; Schmitt, Grandage, & Adolphs, 2004). The implementation of corpus research has revealed that the speakers’ intuitions on their own language are not always accurate, and that methods of introspective analysis often produce descriptions that do not match the reality of textual usage of formulaic sequences (Corpas Pastor, 2003; Sinclair, 1991).

However, Ruiz Gurillo (2001) claims that corpus research on phrasal vocabulary in Spanish is not extensive. Corpsas Pastor (2003) and Ruiz Gurillo (2001) advocate for the use of linguistic corpora in order to investigate diverse aspects of Spanish phraseology such as fixedness, variability, idiomaticity and pragmatic functions of formulaic sequences, and to promote the advance of this linguistic discipline.

---

2 The Real Academia Española (Royal Academy of the Spanish Language) has compiled a 180-million word electronic corpus comprised by written and spoken registers, and constituted by two parts: the
Computer searches are being increasingly performed within spoken and written corpora to locate strings of words, which, according to their recurrent character, i.e., high frequency, may be categorized as formulaic sequences (Moon, 1998; Wray, 2002).

Ruiz Gurillo (1995) explored the frequency of a kind of formulaic sequences within a corpus of colloquial spoken discourse in Spanish, and observed one occurrence per minute of conversation.

Corpas Pastor (2003) investigated the frequency of 100 paremias, i.e., proverbs, selected randomly, in the CVB corpus of peninsular Spanish and found 79 of them in 166 occurrences, with a mean of 2.10. According to the author, these results reveal the high frequency of usage of proverbs in the Spanish discourse (pp. 83-107).

In a study where Spöttl & McCarthy (2002) found evidence that formulaic sequences can be transferred holistically by multilingual individuals across their various languages, the selection of the formulaic expressions accessed was primarily determined on the basis of the high frequency of one of their constituents, generated through corpus analysis. Thus, corpus software can assist in identifying expressions that are potentially formulaic in the opinion of the investigator, for their further examination (Read & Nation, 2004).

CORDE (Corpus Diacrónico del Español), and the CREA (Corpus de Referencia del Español Actual).
This corpus constitutes a valuable source for the study of the Spanish language.
3 Sintagmas prepositivos fraseológicos.
4 The CVB (corpus Vox-Bibliograf) is a ten-million word corpus of peninsular Spanish constituted by a variety of texts published from 1950 to present.
5 The corpus used was the five-million word CANCODE spoken corpus of British English (Cambridge and Nottingham Corpus of Discourse in English).
1.5.2. Problems of corpus analysis

However, Wray (2002) argues that the establishment of the length of multiword expressions and the frequency thresholds is basically arbitrary. Moreover, frequency measures are unable to discriminate between the formulaic and novel nature of identical strings of words (Read & Nation, 2004; Wray, 2002). These subjectivity factors are evident in the enormous disagreement existent on what it is important and how to recognise it (Wray, 2002, p.28). Another proof of this is the number of different frequency and association ratio measures that have been proposed for multi-word expressions explored through corpus research (Read & Nation, 2004; Wray, 2002).

One more problem recognised for these frequency count measures is the size of the corpora (Wray & Perkins, 2000). Small size corpora may not contain, in many cases, some common but less frequent strings of words. In addition, a number of deep-rooted fixed expressions that are undoubtedly familiar to the members of a particular speech community have not been found at all even in mega-corpora studies (Corpas Pastor, 2003; Read & Nation, 2004). Thus, frequency is relative to the corpus one is searching and is not always an essential trait of formulaic language (Corpas Pastor, 2003; Wray & Perkins, 2000).

An additional problem in identifying formulaic sequences in corpora searches relates to the level of variability or flexibility of such sequences. Schmitt & Carter (2004) mention that whereas formulaic expressions constituted by fixed elements could be identified with more ease, flexible sequences, which besides their fixed components have slots that may be filled in with a range of words according to the particular situation, are difficult to uncover using current computer software, even though these expressions may be more frequent than the fixed ones. Similarly, there
are sequences whose elements are located so far away from each other within the discourse that cannot be easily identified by corpus software (Corpas Pastor, 2003).

Schmitt, Grandage & Adolphs (2004) investigated if recurrent clusters identified through corpus research and inserted in a psycholinguistic language task were stored as whole units or not in the mind of the participants. They distinguished between simply recurrent strings of words recognised through corpus analysis, and formulaic sequences, i.e., strings of words that may or may not be located in the corpus search but are stored holistically as formulaic expressions in the mind. They concluded that dependence solely on corpus data would not provide a reliable indication of formulaicity, as recurrent clusters are not necessarily stored holistically in the mind. This is related to the lack of statistical tools that facilitate the semantic analysis of particular strings of words within corpora (Corpas Pastor, 2003).

Corpora examination through computer software constitutes a valuable source of information on various aspects of formulaic sequences in a particular language. However, further manual analyses must be carried out to discard inadequate associations of words that the search tools cannot properly identify, and care should be taken when making intuitive decisions to decide the relevance of the different lexical patterns (Corpas Pastor, 2003; Read & Nation, 2004; Wray, 2002).
1.6. Methodological Issues

1.6.1. Superlemma Theory

This language production theory is central to this study, as the key research instrument used to investigate the participants’ knowledge of certain formulaic sequences was developed on the basis of its propositions.

According to Cutting and Bock (1997), although every idiom has its own individual ‘lexical concept node’ and it is stored holistically on some processing level, idioms have also an internal structure constituted by their compositional simple lemmas that can function within a particular idiomatic framework, but also maintain their individual syntactic and semantic properties in freely produced expressions. When an idiomatic lexical concept node is triggered, the lemmas that constitute the idiom get activated individually, but this activation also spreads to syntactic information in the form of a prefabricated multi-word expression stored as a whole. Moreover, many of the syntactic properties of the idioms have not been found to be different from those of single words (Jackendoff, 1995).

Sprenger, Levelt & Kempen (2006) present an extension of Cutting & Bock (1997)’s hybrid model (Figure 1) with a model that solves the paradox that the holistic nature of idioms, that is, formulaic expressions, is not in conflict with their production by means of single words that retain their own conceptual meanings. This model (Figure 2) introduces a superlemma activation node in speech processing that binds together the constituent lemmas, i.e., single words, of a phrasal lexical item in a single entry in the mental lexicon. The superlemma is a separate “representation of the syntactic properties of the idiom that is connected to its building blocks, the simple lemmas. By representing idioms with their own lemma, idiom production follows the
same rules of lexical competition and lexical selection as single words do” (Sprenger, Levelt, & Kempen, 2006, p.176).

Therefore, the activation of a phrasal item’s lexical concept triggers the activation of its superlemma node, which in turn activates all its constituent lemma nodes that are related to that particular phrasal item’s phonological, phonetic and graphic forms. This allows the respondents to a cloze test to access the missing word in their mental lexicon and produce it (Kuiper, Columbus, & Schmitt, to appear).

**Figure 1**


“Representation of the idiom *hit the road* in terms of the hybrid model, with two types of connections between the concept and the lemma level. The phrasal frame is not depicted.” (Sprenger, Levelt, & Kempen, 2006, p.176)
"Representation of the idiom *hit the road* according to the adapted hybrid model. The idiom is represented both at the concept level and the lemma level. All connections between processing levels denote the same relationship." (Sprenger, Levelt, & Kempen, 2006, p.176)

1.6.2. *Cloze test methodology*

Cloze testing is a common methodology widely used to assess language ability such as reading comprehension skills and second language proficiency (Abraham & Chapelle, 1992; Alderson, 1979; Dörnyei & Katona, 1992; Kobayashi, 2002; Schmitt, Dörnyei, Adolfs, & Durow, 2004). Alderson (1979) argues that some evidence shows that, in general, the cloze procedure works better to test grammar and vocabulary than reading comprehension abilities. It is a written measuring instrument frequently used as a major language testing-tool because it is easy to construct and run, and entails high reliability and validity (Katona & Dörnyei, 1993). This
procedure is defined on a general level as “the systematic deletion of words from text” (Alderson, 1979, p.219) and essentially requires the examinee to fill in the missing words in a sentence where the omitted items are indicated by a blank line. Deletion rate of words can be pseudo-random (it is never totally random\(^6\)); rational, according to a particular criteria depending on what it is being measured; or a deletion process that follows a specific frequency, for instance, the deletion of every fifth word from text (Alderson, 1979), or every fifth to tenth word (Katona & Dörnyei, 1993).

However, researchers have also expressed a number of concerns regarding the cloze procedure since cloze performance can be easily affected by constituent traits of the test such as difficulty determined by the subject matter, nature of deleted words, amount of context provided, deletion ratio and answer mode, among other factors, and scoring methods (Abraham & Chapelle, 1992; Alderson, 1979; Kobayashi, 2002). Scholars have emphasized the need of being aware of these factors when designing, administering and scoring a cloze instrument. Furthermore, caution is needed when interpreting the results of a cloze test to prevent ambiguity on what it is actually being measured (Dörnyei & Katona, 1992). Therefore, some aspects of this methodology were carefully pondered before designing the cloze instrument administered in this research.

**Difficulty of cloze testing.** According to Abraham & Chapelle (1992), an important determinant of cloze item difficulty is the contextual factor, which entails the amount of context required to recall the item. Evidence shows that difficulty is positively correlated with the amount of context required to produce the correct word (Bachman, 1985; Bensoussan & Ramraz, 1984). Kobayashi (2002) also found that

---

\(^6\) The deletion of certain words would cause the loss of context, an important element for test takers to consider when filling a blank.
the smaller the context required for the restoration of an item, the easier this item becomes. According to Abraham & Chapelle (1992), other determinants of difficulty are the number of syllables in the sentence that contains the item, and the incidence of the item in the whole text. They also suggest that the nature of the required response is another difficulty factor: constructed responses are more difficult than selected responses.

**Scoring methods in cloze testing.** The relationship between cloze items attributes and scoring methods has been proved rather complex. Cloze items complexity varies depending on the kind of words deleted. Thus, it is essential to understand the nature of the words being tested in a cloze test in order to envisage with more clarity what each item measures (Kobayashi, 2002). Abraham and Chapelle (1992) argue that an intrinsic factor of difficulty are the basic cognitive processes that test takers require in performing the cloze task, such as the “psycholinguistic processes involved in reading comprehension [and] retrieving vocabulary from long-term memory … on the basis of semantic and syntactic clues” (p. 470). On this foundation, they propose that content words (i.e., verbs, nouns, adjectives, and adverbs) are more difficult to restore than function words (i.e., articles, pronouns, relative pronouns, etc.).

On the other hand, an objective scoring method is needed for the results to be accurate and interpreted in an unbiased manner. Within a theoretical framework of test validity, it is of vital importance to understand the meanings of cloze test scores to justify their interpretation and the actions based on the results (Abraham & Chapelle, 1992). The cloze test procedure can be problematic to mark in many cases unless an ‘exact word scoring’ method is used, which, then again, is hard to justify and
increases the difficulty of the test (Dörnyei & Katona, 1992). Kobayashi (2002)’s findings disclosed variations in the results of cloze tests due to the use of three different scoring methods: (a) The exact word scoring method; (b) The semantically and syntactically acceptable word scoring method; and (c) The semantically acceptable, but syntactically unacceptable word scoring method. The two semantically acceptable-word scoring methods produced higher results’ means, which suggested that high proficiency second language students might unjustly attain low results using the exact-word scoring method on the assumption that greater language proficiency enables students to think of fitting words that, without being the expected ones, make sense in the context.

1.6.3. Superlemma Theory and cloze testing

As regards the suitability of cloze testing as a methodology that makes it possible to investigate whether or not a speaker knows a particular phrasal vocabulary item, it is assumed that the context provided and some of the constituents of the expression can trigger this expression in his/her mental lexicon, allowing the examinee to produce the missing word if he or she is familiar with that formulaic utterance (Kuiper, Columbus, & Schmitt, to appear). This assumption is supported by the superlemma theory which is a model of lexical access during the production of phrasal lexical items (Sprenger, Levelt, & Kempen, 2006). When the subject fails to produce the conventional word that has been deleted from the sentence and replaced by the gap, and simply guesses an alternative filler – which would not necessarily be incorrect or illogical from a syntactic/semantic point of view, then it is believed that this particular formulaic expression has not been acquired and stored in his/her mental lexicon.
1.7. Research Rationale

1.7.1. Antecedents: Acquiring phrasal vocabulary

This research replicated the experimental design carried out in the study *Acquiring phrasal vocabulary* by Kuiper, Columbus, & Schmitt (to appear), which used cloze testing to explore when native speakers of English learn formulaic expressions, what are the differences between native and non-native speakers in their acquisition, and the links between the frequency of usage of the head-verbs of formulaic language items and the mastery of such items by both native and non-native speakers of English.

In the original study, native speakers of English were all New Zealanders assigned to four different groups: 16 ± 1 years of age, between 20 and 30, between 40 and 50, and over 65. Education ranged from high school students to University degrees. Non-native speakers were constituted by two groups: ten German secondary students aged around 16 years of age who had been studying English at a secondary school, and a group of adults over 20 years old who were native speakers of a variety of languages other than English. Neither gender nor socio-economic status were controlled.

The cloze instrument used in the Kuiper, Columbus, & Schmitt’s study was a short story about a social event written in a vernacular style where the participants’ knowledge of twenty formulaic expressions was tested. All test cases where verbal phrases (VP’s). Their verbs, which were gapped in all cases, were classified into three frequency bands and four categories: high frequency light verbs, high frequency lexical verbs, medium frequency lexical verbs and low frequency lexical verbs.
Results suggested that non-native speakers’ acquisition of vernacular formulaic expressions is of a lower magnitude than that of native speakers of English. The means of correct answers for native speakers were 11.9 for adolescents, and 16.4 for adults. On the other hand, non-native speakers’ means of correct answers were 0.8 for adolescents, and 1.8 for adults.

In addition, for native and non-native speakers of English, data showed a more successful retrieval of expressions containing high light and high frequency verbs as compared to lower frequency verb expressions. Results also indicated that there were differences in the number of correct answers among native speakers as a function of age, where older people scored higher than younger people. However, a decline in the number of correct answers was found for the participants over 65 years of age.

Kuiper, Columbus, & Schmitt (to appear) argue that the findings obtained through the cloze testing procedure support previous hypotheses on the acquisition of formulaic language which point out that this process is age graded and that non-native speakers have a lower rate of acquisition. Most importantly, a key result of their investigation is the suggestion that the frequency of the head-verb that is gapped affects the recall of the corresponding formulaic sequence, since expressions containing more frequent verbal items were easier to recall than the ones that incorporated low frequency verbs.

1.7.2. The present study: The acquisition of phrasal vocabulary by non-native speakers of Spanish

The target language of this study was Spanish instead of English. As in the original investigation, a cloze procedure was used to test the following three main hypotheses: a) There are significant differences between the degree of acquisition of
formulaic sequences in native and non-native speakers of Spanish; b) The frequency of usage of the head-verbs contained in verb plus complement formulaic sequences is tightly linked to the acquisition of such sequences; and, c) Phrasal vocabulary of native speakers is age graded in that much of it is acquired in late adolescence and adulthood.

The results of this study were then expected to support the predictions that non-native speakers acquire a significantly less extensive phrasal vocabulary than native users of the language; that the frequency of usage of the Spanish head-verbs in formulaic expressions is tightly linked with the mastery of such utterances; and, that the amount of formulaic language in Spanish acquired is positively correlated with age.

Moreover, available evidence suggests that cultural integration seems to be closely linked to the acquisition of formulaic language (Adolphs & Durow, 2004; Dörnyei, Durow, & Zahran, 2004). Thus, a cultural questionnaire was developed which was intended to measure the participants’ cultural integration level to the target language community to explore its effects on the participants’ acquaintance with certain verb plus complement formulaic expressions.
METHOD

2.1. Participants

There were 94 participants divided in two groups: Fifty five native speakers of Spanish, ages ranging from 15 to 83 years (mean age = 39.01) and 39 non-native speakers of Spanish with an age ranging from 18 to 68 years – one did not specify age, (mean age = 39.32).

The native speakers were all Mexican, and the non-native speakers were either living in Mexico at the time of the completion of the survey, or had lived in Mexico for some time – one did not specify time, ranging from 20 to 600 months (mean time = 165.55 months, i.e. 13.8 years). This guaranteed that the subjects were recruited from a population living in the same language environment, so the formulaic expressions to be explored in the cloze test would very likely have been accessible to all of the participants in the speech produced in a vernacular language in casual and quotidian circumstances in a Mexican community. The native languages of non-native speakers of Spanish were English (20), French (6), German (5), Portuguese (2), Dutch (2), Italian (1), Greek (1), Japanese (1) and Slovak (1).

In some cases Mexican people living in Mexico recruited participants personally. In most cases participants were contacted through the Internet. The persons contacting participants obtained additional information on their background to ensure they fulfilled the necessary requirements. Although educational status was not controlled, university-educated individuals were sought. Some participants, especially the younger ones, were students in High School or University, and most participants had some sort of professional qualification, as per information from the
survey’s administrators. As in the original study (Kuiper, Columbus, & Schmitt, to appear), gender or socio-economic status were not factors considered in the main propositions of this investigation, thus they were not controlled, on the (falsifiable) assumption that they would not be relevant.

2.2. Measures

Two different instruments comprised the survey: a questionnaire intended to measure the participants’ cultural integration level, and a cloze test to explore the participants’ knowledge of certain formulaic expressions. Both instruments were written in the Spanish language and were integrated in a single document which had a cover that contained information, also in Spanish, on the author and on the purpose of the study, some general instructions for the completion of the survey (more specific ones were given at the beginning of each section) and some basic questions in order to get some relevant information from every participant such as their age, native language, and time of learning and/or speaking Spanish in the case of non-native speakers. All questionnaires were anonymous so participants were instructed not to write down their names (see Appendix A for complete survey in Spanish and Appendix B for survey’s cover translated to English).

2.2.1. Cultural questionnaire

The first section of the survey was a questionnaire developed to assess cultural integration of participants to a Spanish-speaking community. It comprised 15 questions that were designed following the model of cultural scales used in some
investigations on second language acquisition (Clement, Dörnyei, & Noels, 1994; Csizér & Dörnyei, 2005; Dörnyei & Clement, 2001; Dörnyei & Csizér, 2002). Answers were given in a 6-point Likert scale format. The total score for each participant was computed on the assumption that the higher the participants’ score, the higher their cultural integration (Minimum score: 15 points; maximum score: 90 points). The statistical reliability of this questionnaire was confirmed through the statistical analysis of its results as it produced a wide variety of scores (see Appendix C for the cultural questionnaire’s items translated to English).

2.2.2. Cloze test

Based on the consideration of the methodological issues of cloze testing mentioned in the introduction section (Abraham & Chapelle, 1992; Alderson, 1979; Bachman, 1985; Bensoussan & Ramraz, 1984; Dörnyei & Katona, 1992; Katona & Dörnyei, 1993; Kobayashi, 2002), and on the superlemma theory on the access and production of phrasal vocabulary (Kuiper, Columbus, & Schmitt, to appear; Sprenger, Levelt, & Kempen, 2006), a cloze instrument was designed for this study to test the understanding and usage of certain colloquial formulaic expressions used in Mexico by native and non-native speakers of Spanish. This took the form of a story called *La Posada*7 about a social occasion written in colloquial language. It was intended to be easy enough to read and provide ample narrative interest in order to stimulate the readers to go through the whole text. *La posada* was designed to provide a sufficient amount of context needed to restore the deleted words, with the aim of offering participants the most appropriate conditions for success in finding the correct word. As in the original experimental design (Kuiper, Columbus, & Schmitt, to appear), this

7 The *posadas* are traditional Mexican parties celebrated daily from the 16th December till Christmas Eve to commemorate (in advance) the birth of Jesus.
cloze test provided participants with significant clues to the meanings of the whole expressions whilst leaving open the possibility of other feasible fillers.

*Deletion ratio.* The deletion process used for this cloze test was rational as the words deleted were the head-verbs of the twenty formulaic expressions selected before creating the test. This decision was also sustained by the suggestion that random deletion should be discarded in favour of rational deletion, supported by a theory of the nature of language and language processing (Alderson, 1979). The objective was to verify if participants did know that particular fixed expression, so they had to produce the correct word that complemented it. Given that the cloze procedure in this study was testing for expressions, not single words, and each of the words sought was a verb that constituted an essential compositional element of the formulaic expression in question, neither the number of syllables in the sentence containing it, nor the incidence of such a word in the text was considered to affect difficulty. The only hint participants could use to access the correct word in their mental lexicon, and then produce it, was the semantic context that bestowed every phrasal item being tested with a particular and exclusive meaning.

*Test items.* Twenty formulaic sequences, i.e., phrasal lexical items, undoubtedly existent in the speech of the community, were selected after an extensive investigation of the context in which they are typically used. For the selection process several phrasal dictionaries were consulted first, and 252 formulaic expressions containing a total of 97 verbs were chosen (Domínguez González, Morera Pérez, & Ortega Ojeda, 1995; Iribarren, 1994; *Locuciones. Diccionario práctico*, 1993;
Sánchez Anaya, 1985; Sánchez Benedito, 1989; Sánchez Benedito & Lavin, 1990; Seco, Andrés, & Ramos, 2004; Varela & Kubarth, 1994).

A second step was to investigate the frequency ranks of the diverse head-verbs contained in such expressions in order to categorise them according to their usage. Extensive investigation on Spanish verb frequency rankings from several sources (Alameda & Cuetos, 1995; Buchanan, 1927; Davies, 2006; Eaton, 1961; García Hoz, 1953; Juillard & Chang-Rodríguez, 1964; Rodríguez Bou & Méndez, 1952) was initially performed. A combined evaluation of the different criteria (frequency, usage, dispersion, merit, range and weighting among others) used in these sources led to the decision of using the information of three main sources: Buchanan (1927), Juillard & Chang-Rodríguez (1964) and Davies (2006). Preference was given to the last frequency dictionary as it contains data obtained from an updated 20,000,000 word corpus where the written sources (fiction and non-fiction) constitute two thirds of the corpus and the spoken sources a full one-third. The other sources’ registers are based entirely on written Spanish from the 1950s or earlier. However, they were still considered as an additional tool to confirm the accuracy of the allocation of the verbs in the particular frequency bands where they had been placed. With slight differences among these sources, the distribution of verbs along the different frequency bands was established to be largely precise. The head verbs of the pre-selected expressions were then classified into four different categories: High Light frequency (or del-lexicalised) verbs (Grimshaw, 1990) (HL), High frequency lexical verbs (H), Medium frequency lexical verbs (M), and Low frequency lexical verbs (L). These categories were defined following a combination of this study researcher’s criteria and the criteria used in the original study (Kuiper, Columbus, & Schmitt, to appear) where
High Light frequency verbs are higher in frequency than the other High frequency verbs.

A final step was the definition of the definitive list of the 20 expressions that were to be included in the instrument, according to the frequency ranking of their head-verbs. Appendix D shows a comparison of the selected verbs frequency rankings in the three sources mentioned above, and Appendix E specifies the frequency criterion. An important consideration that was taken into account when selecting the definitive phrasal vocabulary items was not to include any expression that could be literally translated to English or French to avoid cross-linguistic interference when answering the test, since native speakers of those two languages constituted two thirds of the group of non-native speakers of Spanish. This consideration was based on the fact that there are phraseological units which may totally equivalent in both languages, native and second language, and are used within the same kind of context with the same connotations and rhetorical effects (Corpas Pastor, 2001; García Muruais, 1997). Cloze items that allowed for more than one conventional option (e.g., enter/join the fray) were also excluded. Appendix F presents the final list of formulaic expressions ordered by the frequency of their head-verb.

The selected expressions for the cloze test were presented within the text of La posada with a gap each of which was to be filled by the participant who would try to produce the missing word that would typically complete the formulaic sequence presented. The context presented for each expression was clear and extensive. All 20 formulaic expressions were given in bold type to provide a visual clue that the word to be produced was associated to the bolded succession of words. The missing words
were verbs, all of which belonged to the four different categories mentioned above. Five verbs of each category were included\(^8\).

*Answer method.* Although a multiple-choice methodology would be expected to provide a considerably larger number of correct answers, i.e., recognised phrasal lexical items, the main objective of this research was to verify if participants were able to access from their mental lexicon and then produce the head verbs of the formulaic expressions being investigated, through the activation of the constituent lemmas of such expressions using only the context that was provided in the text as a clue (Kuiper, Columbus, & Schmitt, to appear; Sprenger, Levelt, & Kempen, 2006).

*Scoring method.* For the purpose of this research, the exact-word scoring method was considered to be the appropriate one, as no other words but the precise head-verbs of the formulaic expressions contained in the cloze test would be considered as correct answers. The only way of testing the correct knowledge of the formulaic expressions in question was to look for the exact words that complemented such expressions. Thus, an exact-word scoring method was followed, which added objectivity to the marking process because it was not affected by any kind of subjectivity or personal opinion from the part of the evaluator (Minimum score: 0 points; maximum score: 20 points). Some syntactic variations (i.e., the wrong tense of the verb and spelling mistakes) were ignored since it was the participants’ acquaintance with certain phrasal vocabulary items, and not their grammatical proficiency in Spanish, what was being tested.

\(^8\) The translation of the cloze test *La posada* to English was not considered as relevant since the frequency rankings of the head-verbs of the formulaic expressions selected and the meaning of such formulaic expressions do not correspond to their English counterparts.
2.3. Procedure

Some surveys were administered on a personal basis, but an important part of the data collection was performed by running the survey via Internet, with the support of people living in Mexico in all cases. Helpers were informed about the objectives of the study and were instructed not to give any additional information to participants than the one that was already given on the test itself. They only had to recruit potential participants, make sure participants understood the instructions and the steps to fill the questionnaires in and send them back, and thank them for their help and support.

The email requesting help from potential participants, in the cases where the survey was done through the Internet, also included precise instructions on how to complete the survey on their computer and send it back through the web⁹.

All participants were informed about the importance of giving reliable and honest information for the purposes of the study. No specific information on the objectives of the study was given to participants; they were only told the name of the researcher, the degree she was enrolled for and the name of the academic institution. Such information was specified on the survey’s cover. An important request to all participants was not to ask for help to anyone else when completing the survey, either native speaker or non-native speaker of Spanish. It was an anonymous test and they

---

⁹ Instrucciones para contestar la encuesta:

1. Extrae el archivo a una carpeta de tu máquina.
2. Ejecuta el archivo desde esa carpeta y contesta el cuestionario. Todas las preguntas deberán ser resueltas, aun cuando parezca que no aplican.
3. Una vez que hayas terminado, salva el archivo.
4. Mándame el archivo a este mail __________, extrayéndolo de la carpeta en la que lo guardaste originalmente.

De antemano, te agradezco mucho tu ayuda.
were told that there were no right or wrong answers, so they did not have to worry about their results. Participants were asked to fill in the cover of the survey with some general information and were instructed to ensure they responded to the main two sections of the survey: the cultural questionnaire and the story La posada.

Time of completion of the survey was not controlled although the participants were told to answer questions of the cultural questionnaire and fill the gaps in the cloze test with their first instant thought.

The researcher collected all the surveys that were administered on a personal basis on a trip to Mexico, and printed the surveys that were completed through the Internet, which were sent back to her either by her assistants in Mexico or by the participants themselves.
RESULTS

In this section I will first report results testing the hypothesis that native speakers should have more extensive phrasal vocabulary than non-native speakers of Spanish. Second, I will deal with the critical hypothesis that the mastery of formulaic sequences is linked to the frequency of the usage of head-verbs. Third, I will present results that support the existence of links between age and the mastery of formulaic language for native speakers. Fourth, results that sustain the relationship between time of learning and/or speaking Spanish and the mastery of formulaic language for non-native speakers will be presented. Finally, I will assess the links between cultural integration and the acquisition of formulaic sequences for native and non-native speakers of Spanish.

3.1. Differences between Scores of Native Speakers and Non-Native Speakers of Spanish in the Cloze Test

A preliminary analysis of the means and standard deviations of the scores of the participants showed that, as predicted, native speakers were acquainted with more of the formulaic expressions being tested and achieved higher scores in the cloze test, measured by the number of correct answers, than non-native speakers of Spanish (see Table 1).
Table 1: Means and Standard Deviations for the Scores in the Cloze Test by Both Groups of Participants

<table>
<thead>
<tr>
<th>Origin</th>
<th>N Participants</th>
<th>Mean</th>
<th>Std Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>55</td>
<td>15.2</td>
<td>2.6</td>
</tr>
<tr>
<td>Non-native</td>
<td>39</td>
<td>5.4</td>
<td>4.1</td>
</tr>
</tbody>
</table>

A *t*-test for independent samples was conducted to check if there were significant differences between the scores of native speakers and non-native speakers of Spanish, *t*(60) = 12.905, *p* < .001. As expected, native speakers achieved much higher scores in the cloze test than non-native speakers of Spanish. These results support the prediction that non-native speakers acquire a significantly less extensive phrasal vocabulary than native users of the language.

3.2. Is the Frequency of the Head-Verb of a Formulaic Expression Linked with the Knowledge of Related Expressions?

The results are shown in Table 2 and Figure 3. Because Mauchly’s Test of Sphericity showed that this assumption was violated, Mauchly’s W = 0.812, $\chi^2 = 18.85$, $df = 5$, *p* < .001, a MANOVA analysis was performed. A 2 (native versus non-native) X 4 (head-verb frequency), with a repeated measures on the second factor, MANOVA was thus conducted with the cloze test scores as the dependent variable. On a difficulty scale, expressions containing High Light frequency verbs were considered the easiest ones, followed by expressions with High frequency verbs and
Medium frequency verbs, ending with the expressions including Low frequency verbs, which were considered the hardest ones on the scale.

Consistent with the prior t-test, there was a significant effect of origin (native versus non-native speakers) on the grand total of the cloze test, $F(1,92) = 192.52$, $p < .001$, which showed that native speakers attained higher scores ($M = 15.2$) than the non-native group ($M = 5.4$).

The results also showed a significant effect of verb frequency on the number of correct answers in the cloze test, $\lambda = 0.49$, $F(3, 90) = 30.61$, $p < .001$. As predicted, both groups achieved a higher number of correct answers for the formulaic expressions with High Light frequency verbs than for the formulaic expressions using Low frequency verbs, with the High and Medium frequency being in the middle. Finally, the MANOVA also showed a significant interaction between origin and verb frequency (difficulty), $\lambda = 0.77$, $F(3, 90) = 9.05$, $p < .001$. Thus, the effect of verb frequency on the cloze test total was different for the native speakers than for the non-native speakers of Spanish. As can be seen in Figure 3, this interaction is a function of the non-native speakers largely producing the predicted pattern in which they achieved the highest score on the easiest category, whereas the native speakers produced a flat pattern of results across the easiest three categories.
Table 2: Means and Standard Deviations of the Number of Correct Answers in the Cloze Test for Both Groups of Participants

<table>
<thead>
<tr>
<th>Origin</th>
<th>Frequency</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-native</td>
<td>Low</td>
<td>.51</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>1.59</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>1.31</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>High Light</td>
<td>2.03</td>
<td>1.16</td>
</tr>
<tr>
<td>Native</td>
<td>Low</td>
<td>3.00</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>4.22</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4.16</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>High Light</td>
<td>3.80</td>
<td>.89</td>
</tr>
</tbody>
</table>

NOTE: There were a total of 94 participants: 39 non-native and 55 native speakers of Spanish. Ages ranged from 18 to 68 years old for the first group, and 15 to 83 years for the second group.

Figure 3

Means of the Number of Correct Answers for the Four Categories of Formulaic Expressions According to Their Head-Verb Frequency
Planned comparisons were conducted to test for significant differences on the number of correct answers to the expressions containing High Light frequency verbs and the expressions with Low frequency verbs. As expected, results showed significant differences for native speakers, $F(1, 92) = 22.72$, $p < .001$, and for non-native speakers, $F(1, 92) = 57.61$, $p < .001$, as all participants, independently of their origin (native versus non-native speakers), obtained a considerably higher number of correct answers for the cloze items containing High Light frequency verbs than for those items with Low frequency verbs. The results generally support the proposition that the acquisition of formulaic sequences is tightly linked to the frequency of the usage of the head-verbs contained in such sequences.

3.3. Links between Age and the Mastery of Formulaic Expressions in Spanish for Native Speakers

The correlation between age of native speakers and the number of correct answers in the cloze test is shown in Table 3. As predicted, the number of correct answers in the cloze test increased significantly with age, which suggests that older people know considerably more phrasal vocabulary items than younger people.
Table 3: Correlation Coefficients for Native Speakers of Spanish between the Major Variables

<table>
<thead>
<tr>
<th></th>
<th>High Light</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Cloze Test</th>
<th>Cult Quest</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Light</td>
<td>1</td>
<td>.46**</td>
<td>.16</td>
<td>.31*</td>
<td>.67**</td>
<td>-.05</td>
<td>.19</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>.40**</td>
<td>.40**</td>
<td>.78**</td>
<td>.13</td>
<td>.28*</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>.16</td>
<td>.61**</td>
<td>-.03</td>
<td></td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td></td>
<td>.72**</td>
<td>.20</td>
<td>.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloze Test</td>
<td>1</td>
<td></td>
<td></td>
<td>.10</td>
<td>.31*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cult Quest</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

NOTE: Head-verbs of the formulaic expressions contained in the cloze test were classified under four categories according to their frequency as per research in several sources on word frequency in the Spanish language (see Appendix D).

*p < .05, **p < .01

It was also predicted that although the knowledge of formulaic expressions should increase with age, there should be a decline in older stages of life (> 60 years old). To test this prediction a quadratic function was fitted to the data. However, the associated Beta coefficient for the quadratic effect was not significant, $\beta = -.759$, n.s. Thus, these results did not fully support the prediction that the knowledge of formulaic expressions decline for people in older stages of life.

3.4. Links between Time, Age and Culture, and the Mastery of Formulaic Expressions in Spanish for Non-Native Speakers

As expected, for non-native speakers age was not positively correlated with the number of correct answers in the cloze test (see Table 4) and was close to zero. However,
as expected, those who spent longer learning the language achieved higher levels of language mastery. Thus, age was not an influential factor on the mastery of formulaic expressions for non-native speakers of Spanish, but time spent learning and/or speaking this language is.

Table 4: Correlation Coefficients for Non-Native Speakers of Spanish between the Major Variables

<table>
<thead>
<tr>
<th></th>
<th>High Light</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Cloze Test</th>
<th>Time</th>
<th>Cult Quest</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Light</td>
<td>1</td>
<td>.63**</td>
<td>.60**</td>
<td>.32*</td>
<td>.77**</td>
<td>.41**</td>
<td>.19</td>
<td>.12</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>.79**</td>
<td>.53**</td>
<td>.91**</td>
<td>.38*</td>
<td>.19</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>.50**</td>
<td>.90**</td>
<td>.29</td>
<td>.03</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>.68**</td>
<td>-.08</td>
<td>-.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloze Test</td>
<td>1</td>
<td>.32*</td>
<td>.12</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>1</td>
<td>.38*</td>
<td>.48**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cult Quest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.34*</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The independent variable time was included here as it was hypothesized that for non-native speakers it would be a more reliable predictor for the mastery of formulaic expressions rather than age.
*p < .05. **p < .01.

To test whether time was still significantly related to the number of correct answers in the cloze test after controlling for age and cultural integration, a multiple regression was performed in which the total score was regressed on age, cultural integration and time. Results showed that neither age nor cultural integration were significantly related to language mastery, but that time maintained a significant positive link to language mastery (Beta = .34, p < .05).
3.5. Is Cultural Integration a Significant Predictor for the Mastery of Formulaic Expressions?

The prediction that cultural integration would be positively correlated with the number of correct answers in the cloze test for both groups was not supported (see Tables 3 & 4).
DISCUSSION

“Words do not go together, having first been apart, but rather, belong together, and do not necessarily need separating.”

(Wray, 2002, p.212)

Linguistics and psycholinguistics empirical research has revealed that formulaicity is a central feature of language production and comprehension, and phrasal vocabulary is prevalent in the adult speech of every linguistic community (Corpas Pastor, 2003; Kuiper, Columbus, & Schmitt, to appear; Van Lancker-Sidtis & Rallon, 2004; Wray, 2002; Wray & Perkins, 2000). Evidence shows that the amount of phrasal vocabulary of native speakers is considerably larger than that of non-native speakers (Kuiper, Columbus, & Schmitt, to appear; Wray, 2002). In addition, the results of Kuiper, Columbus, & Schmitt’s (to appear) study suggest the existence of close links between the frequency of usage of the head-verbs of formulaic language items and the mastery of such items by both native and non-native speakers of English. It has also been proposed that formulaicity is age graded. Thus, although the acquisition of formulaic language starts in childhood, much of it is attained in late adolescence and adulthood (Kuiper, Columbus, & Schmitt, to appear; Wray, 2002).

The present research replicated the study Acquiring Phrasal Vocabulary by Kuiper, Columbus, & Schmitt (to appear), to investigate the propositions above, but the target language was Spanish instead of English. Furthermore, given that some evidence points to the existence of a relationship between cultural integration and formulaicity (Adolphs & Durow, 2004; Dörnyei, Durow, & Zahran, 2004), this
research also explored the connection of such an integration and the acquisition of formulaic language.

This section will firstly present a summary of the results, followed by the discussion of the study’s main hypotheses on formulaic language. Next, some cross-cultural implications will be addressed, and then some practical applications of the findings for the pedagogical field of phrasal vocabulary will be suggested. Finally, some strengths and limitations of the present study will be pointed out.

4.1. Summary of the Results

As predicted, native speakers were acquainted with more of the formulaic sequences tested and were able to recall a higher number of the head-verbs gapped in the cloze test than non-native speakers of Spanish. These results support the prediction that non-native speakers of a particular language acquire a significantly less extensive phrasal vocabulary than native users of the language.

Moreover, this study’s findings generally support the prediction that the frequency of usage of the head-verbs contained in formulaic language items is closely linked to the mastery of such items, given that all participants, both native and non-native speakers, achieved a significantly higher number of correct answers in the cloze test for the formulaic expressions containing high light frequency verbs than for the expressions that included low frequency verbs.

Besides, as in the original experimental design, the results of this study support the prediction that the amount of formulaic language acquired by native speakers is positively correlated with age, that is, older people know considerably
more phrasal vocabulary items than younger people. However, age was not an influential factor on the mastery of formulaic expressions for non-native speakers of Spanish, but time spent learning and/or speaking this language was. In effect this may be the same finding since older native speakers have had longer speaking the language than younger speakers.

Finally, although some evidence suggests that cultural integration has a positive effect on the acquisition of formulaic language, there was no correlation found in this study between the score of the cultural questionnaire designed to measure the participants’ cultural integration level to the target language community and the number of correct answers in the cloze test, for either native or non-native speakers of Spanish.

4.2. The Research Hypotheses

4.2.1. There are significant differences between the degree of acquisition of formulaic sequences in native and non-native speakers of Spanish

There is ample evidence of the great capacity of language users to store in the mental lexicon a vast number of fixed expressions in their native language (Jackendoff, 1995; Kuiper, Columbus, & Schmitt, to appear; Mel'čuk, 1995; Van Lancker-Sidtis & Rallon, 2004; Wray, 2002). “And there are parallels in any language” (Jackendoff, 1995, p.135). Whereas this seems to be true for native speakers, important evidence reveals that non-native speakers have great difficulties in the acquisition of formulaic language, and that even proficient foreign language or second language learners store in their mental lexicon a considerably smaller amount
of phrasal vocabulary than native speakers of the language (Bishop, 2004; Kuiper, Columbus, & Schmitt, to appear; Pawley & Syder, 1983; Van Lancker-Sidtis, 2003; Wray, 2000, 2002). Thus, although non-native speakers may be able to acquire over time a large number of formulaic sequences in their second language, they exhibit, in general, deficiencies in their intuitive use and fluency of formulaic utterances, in comparison with native speakers (Kuiper, Columbus, & Schmitt, to appear).

The results of the present study fully support these propositions, given that native speakers achieved significantly higher scores in the cloze test than non-native speakers of Spanish. Native speakers of Spanish knew a large number of the twenty verb plus complement formulaic expressions that were being tested, and were able to recall many of the head-verbs of these expressions that were gapped in the cloze test. The mean of the total of correct answers for native speakers was particularly high (15.2), representing their acquaintance with over three quarters of the formulaic language items in the cloze test, whereas the mean of correct answers for non-native speakers revealed the latter’s knowledge of only one fourth of these items (5.4). A t-test for independent samples confirmed the significant differences between the scores of native speakers and non-native speakers of Spanish.

These results are parallel to those of Kuiper, Columbus, & Schmitt’s (to appear), where native speakers had much higher scores than non-native speakers of English in the cloze test: the means of correct answers for the first group were 11.9 for adolescents, and 16.4 for adults, whereas for the non-native group the means of correct answers were 0.8 for adolescents, and 1.8 for adults. In their study, however, there was a notably larger gap between the scores of both origin groups. A possible explanation may be related to the differences in the length of time that non-native participants in both studies had spent among the target language community, believed
to be considerably larger for the non-native speakers in this study. However, this is merely a conjecture, as the time factor was not controlled in the original study. The time variable will be addressed in more detail later in this chapter.

The results of this study seem to be particularly significant when taking into account the ages of the 55 native speakers that participated in this study, which ranged from 15 to 83 years, because, even though the age variable was an influential factor on formulaic language acquisition, as it shall be seen later, this ample age range did not affect the native speakers’ remarkable performance in comparison to that of the 39 non-native speakers of Spanish, whose ages varied between 18 and 68 years. It must be noted, however, that the outstanding average of the results obtained by native speakers of all ages may be due to the vernacular character of the story that constituted the cloze test, where the sequences being tested were non-specialised lexical items, likely to be known to individuals within a wide range of age.

It has also been proposed that the deficiencies of formulaicity observed in non-native speakers in written contexts, become larger in spoken contexts, as they give preference to the use of single words to their semantically equivalent multi-word expressions (Siyanova & Schmitt, 2007). Thus, it would be very useful to implement future research that explores formulaicity in both written and spoken contexts for the same participants in a study, to test for the differences of performance within and between individuals, that is, within the participants in each origin group, and between both groups of native and non-native speakers.

The lack of formulaicity in the discourse is supposed to pose a relevant problem for the socialisation processes of second language learners within the target language community, given that formulaic language plays an essential role in central functions of social interaction (Schmitt & Carter, 2004; Wray, 2002; Wray & Perkins,
2000). However, from the point of view of some of the administrators of this research’s survey, many non-native participants in the study do not seem to experience this kind of trouble because they have been observed to interact with the other members of the language community, and carry out their personal and work activities without any apparent impediments, compensating with their language proficiency for their lack of formulaicity (evident in their score in the cloze test). These are obviously subjective appreciations but are not that surprising as it is known that language functions in social interaction can also be achieved using novel language (Wray, 2002). Thus, language proficiency may, in many cases, effectively compensate for the absence of native-like proficiency in formulaic language.

4.2.2. The frequency of usage of the head-verbs contained in verb plus complement formulaic sequences is tightly linked to the acquisition of such sequences

The size of the single-word vocabulary does not seem to be correlated with the amount of phrasal vocabulary that a second language speaker possesses (Kuiper, Columbus, & Schmitt, to appear), nor is it a strong predictor of the ability for acquiring formulaic sequences (Schmitt, Dörnyei, Adolphs, & Durow, 2004). Topics such as language aptitude, language motivation, language attitudes, sociocultural adaptation and learning strategies, among others, have been explored in the search of facilitators for the acquisition of formulaic expressions by non-native speakers of a language, without having found strong evidence that points conclusively to any particular factors that assist this acquisition (Kuiper, Columbus, & Schmitt, to appear). But the answer, or an important part of it, is likely to lie within the structure of formulaicity: the characteristics of its constituent words. Could the frequency of usage of the words contained in formulaic sequences be related to the acquisition of
such sequences? And if this were the case, would this fact provide useful insights into
the processes that facilitate the learning of formulaic language?

Kuiper, Columbus, & Schmitt (to appear) originally posed the first of the two
questions above regarding the English language, and used cloze procedure to test
these links. Their findings revealed that the expressions containing high frequency
verbal items were easier to recall than the expressions with low frequency verbs. The
present research constituted in essence a replication of the former, and one of its main
purposes was to investigate if this pattern of verbal frequency could be observed in a
language other than English, in this case Spanish.

As in the original investigation, the results of this study generally supported
the prediction that the frequency of the head-verbs is tightly linked to the
acquaintance with the expressions that contain them, as all participants, independently
of their origin, achieved a significantly higher number of correct answers for the cloze
items including High Light frequency verbs than for those ones with Low frequency
verbs. In Table 2 it can be observed that native speakers of Spanish had a mean of
correct answers of 3.8 for expressions with High Light frequency verbs, and a mean
of 3.00 for expressions with Low frequency verbs, whereas non-native speakers of
this language had a mean of 2.03 for High Light frequency verb-expressions, and a
mean of .51 for expressions with Low frequency verbs\(^\text{10}\). Additional planned
comparisons analyses revealed that these differences were, indeed, significant for both
groups of participants.

Noticeably, the gap between the means of correct answers of High Light and
Low frequency verb expressions was much larger for non-native speakers than for

\(^\text{10}\) Recall that the cloze test presented twenty formulaic expressions, categorised into four groups of five
expressions each, according to the frequency of their head-verbs: High Light, High, Medium and Low.
native speakers of Spanish, generating a ratio of .25 to 1 for the former versus .79 to 1 for the latter. Coincidentally, when calculating the ratio of the total score in the cloze test to the number of formulaic expressions being tested for both groups of participants, the obtained figures were strikingly similar: .27 to 1 versus .76 to 1 for non-native speakers and native speakers respectively. Although no correlation is acknowledged between both types of ratios, these numbers reflect the significant differences existent in the mastery of formulaic expressions between native and non-native speakers of a language.

However, an unexpected pattern was observed for the expressions containing verbs belonging to the High and Medium frequency categories, given that they did not adhere to the verb frequency continuum that had been predicted, which was a linear ascending pattern starting from the lowest point corresponding to the mean of correct answers for the expressions with Low frequency verbs, considered as the hardest ones on a difficulty scale, then going up to Medium frequency, and then again to High frequency, ending up in the highest point corresponding to the expressions including High Light frequency verbs, the easiest ones on the scale.

As it can be seen in Figure 3, the pattern for non-native speakers mostly followed this course, even though the participants scored higher on Medium frequency verbal items than on High frequency ones, contrary to what it was expected. On the other hand, native speakers produced an incongruent flat pattern across the three easiest categories, scoring higher on Medium frequency, followed by High frequency, and then High Light frequency verbal items at a lower position, inconsistently with the predictions. Nevertheless, it is important to note that the differences of the means in these three categories are marginal, given that native speakers achieved a considerably larger number of correct answers for the expressions
in all these three frequency groups, producing the ceiling effect that is observed in
Figure 3.

A plausible explanation of this effect may be found taking a closer look at the
individual items. An unpublished recent work prepared by a postgraduate student of
the Department of Linguistics at the University of Canterbury (Brandt, 2008), which
carried out a similar procedure to the one used in the Kuiper, Columbus, & Schmitt’s
(to appear) study, with a small sample of university students between the age of 19
and 26 years, all native speakers of German, found a similar incongruence on the verb
frequency continuum. The author inspected some of the individual items producing
this pattern, and concluded that cross-linguistic influence and the alternative
selectional preferences for some of the expressions had caused these unexpected
results. When these items were taken out of the graph, the pattern generated looked
much more as predicted.

A thorough investigation of the individual items to be tested in this study was
conducted before designing the cloze instrument. However, the results showed that
two expressions containing High Light frequency verbs were incorrectly answered in
many cases, independently of the origin of the participants. Through a post hoc
informal analysis of these expressions, it could be noticed that one of them, the
expression tener ángel, encompasses a collocate, ángel\textsuperscript{11}, that evokes a range of
verbal alternatives given the context provided. Thus, although the verb tener\textsuperscript{12} is
among the more frequent words used in Spanish, most participants did not recall it
due to the indistinct character of the expression within the context of the story. The

\textsuperscript{11} Ángel means ‘angel’ in English.

\textsuperscript{12} Tener means ‘to have’, but this verb is not used as an auxiliary verb in Spanish in the same way it is
used in English for some verbal tenses.
other expression, *volver* la vista atrás, contains a verb with multiple meanings, a fact that could have caused confusion when trying to recall the correct word. These two expressions generated a significant decrease in the means of correct answers for the High Light category that was reflected in the final results. Yet, it is not possible to reach valid explanations through a merely informal examination of the expressions involved, and a more comprehensive study of the characteristics of the expressions to be included in a cloze test is recommended for future research on formulaicity. One approach is to conduct careful corpus-based investigation on each item such as that conducted by Fellbaum and her associates (2007).

Another explanation may derive from the semantic attributes of the verbs in question. Sinclair argues that there is a tendency to progressive delexicalisation for frequent words. “The more frequent a word is, the less independent meaning it has, because it is likely to be acting in conjunction with other words, making useful structures or contributing to familiar idiomatic phrases” (Sinclair, 1991, p.113). Kuiper, Columbus, & Schmitt (to appear) declare that high frequency verbs such as *make, look,* and *do,* constitute an essential characteristic of informal spoken discourse and are used in innumerable formulaic expressions. Altenberg & Granger (2001), cited in the former study, found that foreign language learners have great difficulty using formulaic sequences that contain the verb *make,* particularly the delexicalised uses. Thus, High Light frequency verbs are more likely to be known by the speakers of a language, but it is precisely this high ‘delexicalising’ frequency that lessens their semantic weight, allowing the legitimate use of other variations, which may seem equally, or even more, plausible.

---

13 To turn, to turn over, to turn inside out; to return, to come or go back; to repay, to restore; to cause, to make; *volver a:* to do again.
Generally, the overall pattern obtained from the examination of twenty formulaic expressions according to the frequency of their head-verbs corresponds to the pattern predicted: High Light frequency verbal items were easier to recall than Low frequency ones. This pattern was more accurate for non-native speakers of Spanish than for native speakers of this language, given the significant differences between the two groups, the latter obtaining a considerably larger number of correct answers in the cloze test in the three highest frequency categories, producing a ceiling effect.

A final reflection on the issue of frequency and on the reason why non-native participants generated a more accurate pattern, i.e., closer to the one predicted, is that if non-native speakers feel more confident by concentrating more on single words than on strings of words when learning and using their second language (Kuiper, Columbus, & Schmitt, to appear; Siyanova & Schmitt, 2007), it would seem sensible that they remember better sequences that contain well known single words they have encountered previously and learned by virtue of their high frequency. However, the question whether non-native speakers store these sequences as ready-made wholes or make them up analytically still remains unanswered and should be addressed in further research.

In any case, if a link does exist between the frequency of the head-verbs of formulaic expressions and the mastery of such sequences, it would greatly benefit the search for strategies that facilitate a second language learner the use of formulaic sequences, which ‘make a non-native speaker sound like a native speaker’.
4.2.3. *Phrasal vocabulary of native speakers is age graded in that much of it is acquired in late adolescence and adulthood*

An integrated model of language acquisition proposes the coexistence of two strategies for processing language: holistic and analytical. The holistic process of language, which is reflected by the use of formulaic language, economises effort and provides the speakers with the fluency they require to be heard. On the other hand, the analytical process brings forth combinations of single lexical items using the rules of grammar (Sinclair, 1987; Van Lancker-Sidtis & Rallon, 2004; Wray & Perkins, 2000).

Research on formulaic language has been profuse in the last few decades, and important findings have uncovered many of its key aspects such as its prevalence in the daily speech of adult native speakers. Although some estimates suggest that the number of formulaic sequences stored in the mental lexicon of a typical native speaker is likely to be ten times larger than the single word lexicon (Mel'čuk, 1995), an accurate approximation of the formulaic component in the speakers internal dictionary, or in a particular language, would be extremely hard to obtain. As Kuiper, Columbus, & Schmitt (to appear) assert, there is not an established list of phrasal items, and, given the ratio mentioned above, along with the inevitable variation of formulaic expressions, such a list would be potentially endless: “We simply do not yet know how many phrasal items there are to know”.

In addition, there is ample evidence which reveals that phrasal vocabulary also plays an important role in the discourse of very young native speakers, that is, children. Therefore, the question is not only about how much, but how and when people learn formulaic language.
Even though some research has been done with the aim of answering these questions, the processes by which speakers originally acquire formulaic language have not yet been fully explained (Kuiper, Columbus, & Schmitt, to appear). However, a generally accepted developmental model has been proposed to describe the different phases that native speakers go through while dealing with language, either holistically and/or analytically, since childhood (Wray, 2002; Wray & Perkins, 2000). This model basically describes the interaction of both strategies during four phases of language development, but emphasises the increasing and prominent nature of formulaic language towards late adolescence, and then adulthood, when its patterns are established. Thus, formulaic language is present in the speech of speakers since infancy, but the amount of phrasal vocabulary that speakers store in the mental lexicon increases significantly with age.

Kuiper, Columbus, & Schmitt (to appear) explored the association of age with the mastery of formulaic sequences for native speakers of English, and their results showed that there is an important effect of age, as older people scored higher than younger people in the cloze test. They found, however, a decline in the number of correct answers for the participants over 65 years of age.

The results of the present study supported most of Kuiper, Columbus, & Schmitt’s outcomes, with the variant that the participants were native speakers of Spanish. As expected, a significant correlation was found between the total score of the cloze test and the age of the participants, given that the number of correct answers increased considerably with age. Thus, the findings revealed that older people knew significantly more phrasal vocabulary items than younger people. On the other hand, the prediction that there might be a decline for older people was not supported. A possible explanation for this result may be that the size of the sample of people aged
over 65 years was relatively small in proportion to the total sample of participants (5 participants out of 55 native speakers, i.e., 9% of the total sample), hence, not properly representing this population and generating a statistical problem of power that may account for this result.

Supposing therefore that formulaicity in native speakers is age graded, how is this to be explained? It is not simply a product of an accumulative process that results from the time of exposure to formulaic language in everyday activities, throughout one’s life. Although time is unquestionably an influential factor for the acquisition of formulaic sequences, there is another central reason for the prevalence of formulaic expressions in the discourse of adolescent and adult native speakers of a language: their social interactional needs.

The use of formulaic language not only increases the efficiency of the language processes, and provides the speakers with the necessary fluency by reducing the time of retrieval of language items from the mental lexicon; formulaicity also performs important language functions in social interaction, and presents the speakers with suitable frames for communicating (Corpas Pastor, 1996; Kuiper, 1996; Nattinger & DeCarrio, 1992; Sinclair, 1987; Wray, 2002).

During childhood, speakers live inside of a ‘socio-interactional bubble’ (Wray, 2002, p.135), where their basic communicative needs of survival and comfort can be fulfilled by using just a few formulaic expressions. Consequently, children can afford the time to develop their analytic language ability, as their verbal exchanges are rather constant across the limited range of the situational contexts where they interact. However, mental and physical maturational processes take their course, and speakers face new communication challenges when encountering new situations and assuming new social roles. Thus, when these times come, adolescents and adults’ language must
be apt to execute the social interaction functions required to guarantee speakers’ adequate integration into the social community where they belong. Formulaic language largely achieves this task. Therefore, the amount of phrasal vocabulary of native speakers increases with age, and it should be expected that this amount be positively correlated with the number and nature of social roles that an individual performs in a particular society.

4.2.4. *Cultural integration is closely linked to the acquisition of formulaic language*

Even though a number of investigations on the connection of cultural integration and language acquisition have been carried out, specific research on the association of such an integration and formulaic language has been meagre in relation to its importance, and the results are contradictory (Adolphs & Durow, 2004; Dörnyei, Durow, & Zahran, 2004; Schmitt, Dörnyei, Adolphs, & Durow, 2004; Siyanova & Schmitt, 2007; Wong-Fillmore, 1976). Besides, much of this research has focused on school-based language learners who often are inclined to adopt an analytical approach to learning, given that the type of language instruction they receive is, most likely, based on single-word vocabulary and grammar, a traditional approach in many schools (Rod Ellis, 2005).

However, the task of defining the term ‘cultural integration’ may not be easy, as it involves a great deal of subjectivity. Gardner (2001) asserts that cultural ‘integrativeness’ reflects “a general interest in learning the second language in order to come closer to the other language community” (p. 5). On the other hand, Dörnyei & Csizér (2002) believe that the term ‘integrativeness’ may not refer to “any actual,
or metaphorical, *integration* into an L2\textsuperscript{14} community as to some more basic *identification process* within the individual’s *self-concept*” (p. 453). In any case, integrativeness has proven to be an important factor in language proficiency by second language learners. Therefore, following the model of some language attitudes and motivation scales used in some important investigations on the acquisition of language (Clement, Dörnyei, & Noels, 1994; Csizér & Dörnyei, 2005; Dörnyei & Clement, 2001; Dörnyei & Csizér, 2002), a questionnaire was developed intended to assess the cultural integration of the non-native participants in this study to the Spanish speaking community.

Contrary to predictions, the results did not show any positive correlation between the scores in the cultural questionnaire and the number of correct answers in the cloze test. These results may not be surprising if it is taken into account that language proficiency, which might be linked to cultural integration, does not guarantee the mastery of formulaic language. It has been observed that adult non-native speakers rely more on the analytic language system and concentrate more on individual words than on sequences of words (Kuiper, Columbus, & Schmitt, to appear). As Wray (2002) indicates, proficient second language learners are able to produce perfect grammatical utterances, which, in spite of their flawlessness, native speakers would never use. Thus, many non-native speakers in this study may well have had positive attitudes to the target language community and been skilful speakers of Spanish, but not necessarily been acquainted with the formulaic sequences tested.

However, there is another important aspect to be considered. Cultural integration, as measured by the questionnaire which was administered, only

\textsuperscript{14} Second language.
implicated the non-native speakers’ attitudes toward their second language and their interest to get closer to the native community, but did not reflect their interaction with the members of the target language community. Such interaction, if assessed, would have probably showed an effect on the degree of knowledge of the formulaic expressions included in the cloze test by the non-native speakers, given that, as it has been mentioned before, formulaicity plays and important role in fulfilling interactional needs in any society (Wray, 2000; Wray & Perkins, 2000). Therefore, through the active contact with native speakers of the target language community, the exposure of non-native speakers to the use of a range of formulaic expressions may facilitate their learning. As Irujo (1986) declares, “input without interaction is not sufficient for language acquisition” (p. 237).

Ellis (2005) argues that the more exposure to the target language non-native speakers experience, the more and the faster they will learn. He also maintains that interaction in the second language is essential to building up language skills. In support of this, it is important to note that time proved to be an influential variable for the non-native group in the present study, showing a significant correlation with the knowledge of the formulaic sequences being tested by the non-native speakers of Spanish. Brandt’s (2008) findings also revealed that the German university students who had lived abroad in an English-speaking country for three months or more scored significantly higher in the cloze test, doubling-up the number of correct answers for the formulaic expressions tested, compared with the students who had only studied English at university, or just had their secondary school English to rely on.

Thus, it is assumed that the period of time living among the target language community represents amount of interaction and exposure. The non-native participants in this study lived, or had lived, in Mexico at the time of the study
between 20 and 600 months. This range of time did prove to have an effect on the results, as participants that had lived longer in Mexico were able to recall many more of the verbs gapped in the cloze test than participants living in the second language community for a shorter period of time.

Social interaction is undoubtedly central to formulaic language learning, and further research on its role on the acquisition of phrasal vocabulary by non-native speakers of a language should assist to clarify many questions in the study of formulaicity in second language learning.

4.3. Cross-Cultural Implications

“Whatever the precise relation of language and thought, it is undeniable that human existence is deeply affected by the ability to speak and understand language.”

(Jackendoff, 1994, p.5)

Corpas Pastor (2003) argues that the emergence of phraseology, the academic discipline that studies the multi-word expressions of a language, i.e., formulaic sequences, goes back to the late 70’s. She distinguishes three main blocs in the study of formulaicity: the Linguistics schools of Eastern Europe, Western Europe and North America. The criteria applied to the investigation of formulaicity by these scholarly approaches may differ in some aspects, but the basic definitional principle is the same: relatively stable sequences of words that are stored holistically in the mental lexicon of language users.
Thus, formulaicity has been explored worldwide for several decades now, showing the undeniable presence of this phenomenon in a variety of languages. Formulaic expressions have been found in a variety of languages, besides English and Spanish, such as Russian, French, Italian, German, Swedish, Polish, Arabic, Hebrew, Turkish, Greek, and Chinese. Kuiper, Columbus, & Schmitt (to appear) note that although this sole fact does not prove its universal character, ‘it strongly suggests that [formulaicity] is a common phenomenon’.

Much of Corpas Pastor’s work has concentrated on the study of the similarities and differences of the phraseological universes among languages. She highlights the existence of formulaic lexical items that may be categorised as ‘phraseological universals’, ‘loans’ or ‘inter-linguistic correspondences’ because of their structural and functional resemblance across languages. She proposes as well the incidence of formulaic expressions that are specific to a particular language or culture and do not have an equivalent in other languages, with a vast number of other expressions occurring in the middle of this continuum. Therefore, the degree of equivalence for the formulaic phrases in different languages may be total, partial or null (Corpas Pastor, 2001).

The cross-cultural character of formulaicity is observed in some studies of formulaicity that reveal the importance of cross-linguistic influence, such as Spöttl and McCarthy’s (2002) investigation where their multilingual participants were able to transfer holistically formulaic expressions across their various languages. Cross-linguistic influence may be the result of common historical or cultural roots, or of the human tendency to formulate metaphorical expressions based on the same general topics (Zalasiewicz, 1991, cited in Kuiper et al., to appear). However, cross-linguistic influence may also interfere with the comprehension of certain formulaic
expressions that include single lexical items that are associated to totally dissimilar concepts in different cultures (Mansilla & Mena, 1997).

Corpas Pastor (2003) has carried out a number of contrastive studies of formulaicity in Spanish and English, and has concluded that it is possible to establish a set of general criteria for the classification and study of the formulaic expressions that is valid and works for these ‘two genetically distant languages’.

The results of this research support this assumption, as they showed a remarkable similarity to those obtained by Kuiper, Columbus, & Schmitt (to appear), pointing to the possibility that the processes of acquisition of formulaic language in Spanish and English function in a very similar way. The findings of both studies revealed that in both languages the extent of phrasal vocabulary of native speakers is significantly larger than that of non-native speakers, that the frequency of usage of the head-verbs of verb plus complement formulaic expressions is linked to the mastery of such expressions, and that formulaicity is age graded.

This constitutes significant evidence of the similarity of some important patterns of formulaicity in both languages, English and Spanish, and the cross-cultural implications of this fact are considerable. If the patterns of formulaic language that are analogous in these two languages are comprehended, the processes by which people acquire formulaic sequences would be easier to understand, and second language learners, either of English or Spanish, would benefit greatly from the development of appropriate pedagogical methodologies and tools that facilitate not only their grammatical proficiency in their second language, but also their native-like use of formulaic expressions.
4.4. Applications

4.4.1. Teaching formulaic language

Language acquisition is frequently referred to as the learning of rules (Wray & Bloomer, 2006). Grammar rules and grammatical analysis of sentences often constitute in second language classroom-based education the fundamental aspects of teaching. Moreover, second language teaching tends to focus on vocabulary learning activities for the acquisition of single words. Therefore, language students usually start by learning individual words, and then put them together using the grammar rules in sequences that, in most cases, are far from sounding native-like.

However, the important findings of the increasing linguistic and psycholinguistic research on the phenomenon of formulaicity have raised awareness of the importance of teaching formulaic sequences to second language learners. The appropriate use of formulaic expressions may not only make non-native speakers ‘sound’ like native speakers; the functional characteristics of the formulaic sequences selected under suitable cultural frames will also ensure that the speakers will be understood. Thus, formulaicity in language promotes the adequate integration of non-native speakers into the particular community where they interact.

Based on the importance of acquiring formulaicity by second language learners, some pedagogical approaches have been developed. Nattinger & DeCarrico (1992), stress the socio-interactional functions of formulaic sequences, and suggest that even though lexical phrases are preconstructed chunks of language reached by the speaker from the mental lexicon, they can mostly be analysed by using the rules of grammar. They argue that language learners can generalise and learn broader grammatical and morphological aspects of their second language by ‘chunking’ and
analysing the formulaic sequences with which they are presented as primary input in their learning process. In a similar line, Lewis (1993) claims that “language consists of grammaticalised lexis, not lexicalised grammar” (p. vi), and also believes in the acquisition of grammar rules through the generalisation from the formulaic expressions learned. Willis (1990) emphasises the existence of recurrent patterns of words in language, and the importance of learning and analysing such patterns by second language students in order to extend their language abilities.

A significant critique to the approaches outlined above is that of Wray’s (2000), who finds it highly contradictory that “in order to encourage the development of native-like idiomaticity, a fundamentally analytic approach is promoted, even though the very nature of formulaic sequences seems to be that they are not normally analysed” (p. 484). Other teaching approaches attempt to find a balance between the analytical and holistic features of the language. Ellis (2005), for example, argues that proficiency in a second language is acquired by the combination of the knowledge of grammatical rules with a vast repertoire of formulaic expressions.

The literature thus shows that the difficulty that non-native speakers have in learning and using formulaic expressions has not been ignored, but as long as the mechanisms by which speakers acquire formulaic sequences are not elucidated completely, it will not be possible to develop an appropriate methodology for teaching phrasal vocabulary. In addition, as Kuiper, Columbus, & Schmitt (to appear) note, there is another important issue to be considered: how to select the most useful formulaic expressions to teach and learn.

The present study’s findings, in agreement with those of Kuiper, Columbus, & Schmitt, suggest that the frequency of usage of the head-verbs contained in formulaic expressions is linked to the mastery of such expressions. If this were the case, this fact
could provide valuable insights not only into the mechanisms of acquisition of phrasal vocabulary by native and non-native speakers of a language, but also into the selection of the most appropriate phrasal lexical items to teach in second language courses, or at least those most likely to be acquired.

4.4.2. The importance of word frequency in vocabulary teaching

Word frequency has a recent tradition of being used as useful criteria in teaching English vocabulary, as corpus research has revealed significant information about the coverage of the different words in written and spoken corpora:

“The word the accounts for 7% of the running words in written texts. The most frequent 10 words account for around 25% of the running words in spoken and written use. The most frequent 1000 words account for around 75% of the running words in formal written texts and around 84% of informal spoken use. By contrast, the tenth 1000 most frequent words account for much less than 1% of the running words in a text” (Laufer & Nation, 1999, p.35).

This information has led to the proposition that, in general, low frequency words are not as useful to teach as high frequency words, given that the ‘return for learning’ the high frequency words, which is the coverage of spoken or written texts that the knowledge of these words provides, is very large (Nation, 1993). Thus, under this proposal, the answer to the question ‘How much vocabulary does a second language learner need?’ rests on the appropriate selection of a set of words that, according to their high frequency of usage, will provide a wide range of coverage of the language that non-native speakers must know in order to successfully understand and communicate in their second language. Therefore, vocabulary teaching should
concentrate on teaching high frequency words, also taking into consideration their semantic and syntactic usefulness, e.g., words that “have many meanings, can define other words, form opposites, have many collocates” (Nation, 2001, p.178), etc.

The majority of the research on the importance of frequency on vocabulary teaching has been done for single-word lexical items. However, it has been suggested that the frequency of occurrence of formulaic expressions, ‘and of the meanings and functions to which they are attached’, also constitutes very useful criteria for the selection of the phrasal vocabulary which is most helpful to teach and learn (Kuiper, Columbus, & Schmitt, to appear). The problem of deciding what phrasal lexical units should be included in language curricula would be simplified through the initial identification of their key elements. Thus, if formulaic expressions containing high frequency verbs are better known to both native and non-native speakers of a language than sequences including low frequency verbs, the selection should focus on the former as they seem to be more prevalent in the discourse and, most importantly, easier to acquire and store in the mental lexicon. In line with this assumption, Willis (1990) affirms: “What emerges very strongly once one looks at natural language, is the way the commonest words in the language occur with the commonest patterns” (p. vi).

However, it is important to remember that the results of the present study did not produce a linear correlation between the frequency rankings of the verbs explored and the knowledge of the formulaic expressions that included them. The findings revealed a broader band of frequency where the three highest categories, High Light, High and Medium, achieved the top results. This points to the fact that, for the creation of a frequency list and the determination of the cut-off points, it is fundamental to take into account a number of other characteristics such as
representativeness, range of usage across language texts and collocational properties, among others, of the verbs in question (Waring & Nation, 1997).

Thus, the exploration of formulaic sequences through corpus research may be approached from various perspectives: the incidence of a particular expression in relation to the frequency of its head-verb, the frequency of different formulaic sequences where a particular verb occurs, the occurrence of formulaic expressions containing high frequency verbs as compared to that of sequences including low frequency verbs, etc. The availability of sophisticated computer programs and large corpora in both languages, Spanish and English, makes possible the investigation of numerous alternatives in the search of an appropriate body of phrasal lexical items to be included in the syllabus of a language course.

4.5. Strengths and Limitations

The greatest strength of this study is that it is the first one that has investigated the connection between the frequency of usage of the head-verbs of formulaic sequences in Spanish and the mastery of such sequences. This study replicated the investigation of Kuiper, Columbus, & Schmitt (to appear) which originally explored this type of relationship, but for English phrasal lexical items. Thus, this study extends the current knowledge of the role of frequency on the acquisition of phrasal vocabulary.

Another strength of this investigation is the cloze instrument designed, which proved to be very effective to test the participants’ acquaintance with some formulaic expressions, by gathering in every case all the answers (correct or incorrect) required
from the respondents. Its vernacular character and light and easy-reading nature encouraged the participants to go through the whole story and provided the context required for all the gaps to be filled in with a word.

However, some limitations should be noted. The cultural questionnaire created to test the cultural integration of the participants did not provide the expected results, as it did not include measures of interaction and qualitative exposure to the target language community, and was limited to assessing integrativeness, basically conceived of as the learners’ attitude towards the second language community.

Additionally, ample evidence reveals that there exist systematic differences between male and female learners in their disposition toward language and their linguistic abilities (Dörnyei & Csizér, 2002; Holmes & Meyerhoff, 2003). Therefore, it would have been interesting to record the gender of the participants in the present study to test for its possible effect on the results concerning the differences in the acquisition of formulaic expressions between native and non-native speakers of Spanish.

Finally, an important limitation was the impossibility of applying the survey in situ, which prevented the control of some significant factors, such as the time that respondents took to answer the questionnaires, and the chance that participants sought help from outside sources.
CONCLUSION

The above limitations notwithstanding, the findings of this study have provided strong support for previous hypotheses on the acquisition of formulaic language which propose that this process is age graded, and that non-native speakers have a lower rate of acquisition than native speakers of a language. Furthermore, a significant result of this investigation is the suggestion that the frequency of usage of the head-verbs contained in verb plus complement formulaic expressions in Spanish is positively correlated to the acquisition of such expressions, given that sequences including high frequency verbs were easier to recall than expressions with low frequency verbs by both groups of participants, native and non-native speakers of this language.

These results are parallel to those obtained by Kuiper, Columbus, & Schmitt (to appear) for the English language, a fact that suggests the existence of cross-cultural implications of the phenomenon of formulaicity. Thus, further research on the identification and comprehension of the analogous patterns of formulaicity among languages may unveil important information regarding the processes of acquisition of formulaic expressions by both native and non-native speakers of a language, in this case, Spanish and English.

Moreover, the investigation of the connection between the frequency of usage of the head-verbs of formulaic sequences and the mastery of these sequences constitutes a new and promising field of study for the understanding of the processes involved in the acquisition, storage and use of formulaic language. Further studies will test if this type of frequency data can assist in the identification and selection of the formulaic expressions that are best to be taught, and easy to be learned, and in the
development of the most appropriate methodologies for teaching phrasal vocabulary. In addition, research in this area will undoubtedly contribute a great deal to the elaboration of comprehensive and useful dictionaries for second language learners.

The acquisition of phrasal vocabulary is an essential part of the language learning process. It is important for language teachers to understand the nature of formulaicity and the linguistic and developmental processes that underlie its usage. Therefore, by having explored and identified some of the important features of formulaic expressions, it is hoped that the outcomes of this study will contribute collaterally to this endeavour, helping in the development of an appropriate pedagogy for teaching phrasal vocabulary in Spanish.
REFERENCES


APPENDICES

Appendix A: Survey in Spanish:
  - Cover
  - Cultural questionnaire
  - Cloze test *La posada*

Appendix B: Cover translated to English

Appendix C: Cultural questionnaire translated to English

Appendix D: Head-verbs frequency rankings

Appendix E: Frequency criterion

Appendix F: Final list of formulaic expressions
Appendix A

Survey in Spanish:
- Cover
- Cultural questionnaire
- Cloze test *La posada*
Muchas gracias por participar en esta investigación llevada a cabo por Victoria Escaïp, estudiante de postgrado de la Facultad de Psicología de la Universidad de Canterbury en Christchurch, Nueva Zelanda. Este no es un examen así es que no hay respuestas correctas o incorrectas y usted no necesita escribir su nombre. Los resultados de esta encuesta serán utilizados solamente para propósitos de investigación por lo que se le pide responder honestamente. La información que usted nos proporcione será vital para el éxito de este proyecto.

Edad __________ (Por favor no escriba su nombre)

¿Cuál es su lengua materna? Español ( ) Inglés ( )
Otra ( ) ¿Cuál? __________

Si su lengua materna no es el español, ¿Cuánto tiempo lleva aprendiendo y/o hablando español? _______ años _________ meses

Esta encuesta contiene dos secciones:

Sección I  Cuestionario cultural

Sección II  La posada

Por favor asegúrese de completar las DOS secciones. Muchas gracias.
**Sección I**

**Cuestionario cultural**

Lea cuidadosamente las siguientes preguntas y encierre en un círculo el número que sea más apropiado para usted. Por favor NO deje ninguna pregunta sin responder. Muchas gracias.

<table>
<thead>
<tr>
<th>Definitivamente no</th>
<th>Muy poco</th>
<th>Poco</th>
<th>Más o menos</th>
<th>Mucho</th>
<th>Muchísimo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

1. ¿Le parece interesante aprender español?  
2. ¿Le parece interesante hablar español?  
3. ¿Encuentra divertido aprender español?  
4. ¿Le gusta hablar español?  
5. ¿Considera que hablar español es importante hoy en día?  
6. ¿Le simpatiza la gente que vive en Latinoamérica y España?  
7. ¿Le gusta hablar español con hispano-parlantes nativos?  
8. ¿Piensa que es importante conocer a gente de países de habla hispana?  
9. ¿Le gusta viajar a países de habla hispana?  
10. ¿Le gustaría saber más sobre la gente de los países de habla hispana?  
11. ¿Le interesan los valores y costumbres de los países de habla hispana?  
12. ¿Le gustaría saber más sobre la cultura de los países hispanos?  
13. ¿Le gusta la música en español?  
14. ¿Le gustan las películas en español?  
15. ¿Le gustan los programas de televisión producidos en los países de habla hispana?
Sección II

Por favor lea la siguiente historia y escriba en los espacios la palabra adecuada en su opinión. Por ejemplo:
“Nunca volví a saber nada de María, le perdi la pista hace muchos años.”

No se preocupe si no tiene seguridad de que sea la palabra adecuada, simplemente escriba lo primero que le venga a la mente.

Contra todos sus deseos, Vanesa pidió un taxi para ir a la posada de la compañía. Realmente no tenía ganas de ir, pero sabía que debía hacerlo pues a pesar de tener ya algunos meses trabajando allí, no conocía a mucha gente y ésta era una buena oportunidad para presentarse y relacionarse con el personal de otras áreas. La noche anterior lo había ______________ con la almohada y había decidido que iría. No le harían mal un par de tequilas y algo de música después de todo.

Esta época navideña le producía nostalgia pues, al _____________ la vista atrás, contemplaba a una niña ilusionada y juguetona que no veía la hora de abrir los regalos de Navidad. Pero el tiempo pasa, uno crece y la sociedad te indica que tienes que ______________ cabeza. La magia de la niñez se evapora. Así es que ahora Vanesa se entristecía de no poder sentir ese espíritu navideño al que mucha gente se ______________ en cuerpo y alma.


Sin embargo, tenía que ser honesta con ella misma… no sólo era su condición de mujer adulta de casi 30 años la que la prevenía de ilusionarse en esta época… Precisamente en diciembre del año anterior, Rodrigo, su novio de la preparatoria, de la universidad y de toda la vida, había decidido ______________ el ala y volar lejos con su mejor amiga. Fue un golpe terrible para Vanesa. Tuvo una depresión profunda, no se levantó de la cama ni probó bocado durante días y lloró por semanas hasta que su madre habló con el tío Pablo, el millonario de la familia dueño de una gran empresa y le suplicó que le ofreciera un trabajo a Vanesa en su sucursal de Querétaro. Eso la sacaría de la ciudad de México y la mantendría alejada de los lugares que le traían recuerdos tan tristes.

De algo le había servido a Vanesa ______________ las pestañas tantos años estudiando Derecho. Había sido la mejor estudiante de su generación y el tío
Pablo la había hecho Subdirectora Jurídica de Equipos Electrónicos Modernos, S.A. de C.V. Todavía no se sobreponía completamente de su pena de amor, pero ahora veía todo aquello desde otra perspectiva. Se preguntaba cómo había podido enamorarse tanto de Rodrigo. Era cierto, Rodrigo ____________ ángel, era guapo y simpático, pero no era una persona de fiar. Nunca cumplía lo que decía, siempre ____________ a su palabra. Además, le ____________ el cerebro a cualquiera con tal de conseguir sus objetivos. ¡Pero qué ciega estaba! ¿Qué tonta, no puedo creer que casi ____________ el suelo que Rodrigo pisaba. Tanto tiempo perdido... ¿Por qué no lo mandé a ____________ espárragos desde el principio...? pensó sintiendo un inesperado calor subiéndole de los pies a la cabeza, pero inmediatamente recuperó la compostura. Quería llegar a la fiesta serena y sonriente, y ya no se iba a ____________ en un vaso de agua por culpa de aquel tipo que no la merecía. ¡Nunca más le iba a ____________ la fiesta!


Vanesa se bajó del taxi y se dirigió ligera en sus altas zapatillas italianas rojo sangre hacia la casona de piedra rosada que tenía frente a ella. Tocó el timbre en la pared junto al portón de madera tallada. Podía escuchar la música, las voces y las risas que provenían de la casa. Luis Alberto, el dueño de la casa, le abrió la puerta. Le sonrió ampliamente, y para no perder la costumbre, le comenzó a ____________ flores. “Bienvenida mi querida Vanesa, tu elegante belleza me alegra la vista. El rojo es definitivamente tu color... ¡¡¡Te ves espectacular!!!” “Gracias Beto, tú tan caballeroso como siempre.” Sonrió Vanesa sacudiendo su melena castaña y sabiendo lo mucho que ella le gustaba. “¡Qué linda tu casa!” Lo decía en serio. Era una antigua casa de estilo colonial, que a pesar de los años se mantenía hermosa e imponente. Luis Alberto pertenecía a una adinerada familia de Querétaro que había pagado por sus estudios de Ingeniería en el MIT de Massachussets. Ahora, apenas en sus treintas, él era el Director Ejecutivo de la sucursal Querétaro de la compañía.

“Pasa, pasa, por favor, llegas a tiempo pues estamos a punto de comenzar la peregrinación.” La tomó de la mano mientras la conducía por el inmenso jardín donde, aún sin haber conocido el interior de la casa, saltaba a la vista la riqueza de los propietarios. Una mano cálida... “Aquí están las velitas. ¿Dónde quieres cantar, adentro o afuera?” Siempre tan amable y considerado pensó Vanesa al tiempo que
prendía su velita blanca y afinaba la garganta para pedir posada. Luis Alberto le
gustaba, no lo podía negar, pero desde su ruptura con Rodrigo había decidido ser más
precavida y no ______________ el tiempo con amoríos superficiales. En todo
caso, ella no iba a ______________ el primer paso.

“En el nombre del cieeeecelo, os pido posaaaada…” comenzó a cantar con el
grupo que se concentraba en el enorme jardín frente a la puerta de cristal que daba
acceso a la casa donde otro grupo cantaba “Aquí no es mesooooón, sigan
adelaaaaante…” Vanesa empezó a relajarse cuando le pareció reconocer a alguien
a través del cristal de la puerta. “¿Quién es ese hombre de saco azul? ” Le preguntó a
Luis Alberto, quien había permanecido a su lado. “No trabaja en la empresa, se
_______________ en la fiesta, ya ves cómo algunas personas lo hacen para poder
_______________ el codo gratis, jajaja. Bueno, en realidad es hijo de unos amigos
de mis padres. Es abogado y está aquí porque no le ha ido muy bien en la ciudad de
México y quiere que yo lo ayude a conseguir trabajo en la empresa. Pensé que tal vez
en tu área haya algo para él. Se llama Rodrigo.”

Vanesa tuvo que morderse la lengua para no gritar su indignación, su rabia, su
despecho… pero, para su sorpresa, estos sentimientos sólo duraron unos cuantos
segundos y como por arte de magia la invadió una gran calma que le permitió
responder imperturbable “No Beto, no hay nada para él. Lo conozco bien y por nada
del mundo lo recomendaría en ningún lugar. Hoy no me preguntas por qué, pero algún
daía te voy a ______________ el cuento.” Vanesa se acercó un poco más a Luis
Alberto, lo tomó de la mano y lo miró coquetamente a los ojos, Mmmh, son verdes…
“¿Me invitas algo de tomar? ” Luis Alberto apretó su mano, le devolvió la sonrisa y la
atrajo un poco más hacia él. “Entendido. Confio ampliamente en tu capacidad y
criterio. No hay trabajo en esta empresa para el hombre del saco azul”, y la encaminó
suavemente hacia el bar.

Vanesa bailó, conversó, rió, bebió y rompió la piñata celebrando el inicio de
esta nueva etapa en su vida. No le ______________ la palabra a Rodrigo en toda
la noche. ¿Qué pasaría con Luis Alberto? No lo sabía, ¡pero lo iba disfrutar mucho!
Ahora era más madura, tenía más confianza en sí misma y sabía lo que quería. ¿Qué
pasaría con Rodrigo? No le importaba. Esta noche, después de mucho tiempo, sentía
como si le hubieran quitado un peso de encima. ¡Nunca habría pensado que la
venganza es tan dulce! Se dijo mientras chupaba la caña de azúcar del ponche con
ron que su futuro nuevo amor de ojos verdes le acababa de traer.
Appendix B

Cover translated to English
APPENDIX B

Thank you very much for agreeing to take part in this survey conducted by Masters graduate student Victoria Escaip from the School of Psychology of the University of Canterbury, NZ. This is not a test so there are no right or wrong answers, and you do not even have to write your name on it. The results of this survey will be used only for research purposes so please answer honestly. Please read each item carefully before responding. The information we get from you will be vital to the success of this project.

Age ________ (Please do not write your name)

What is your native language? Spanish ( ) English ( ) Other ( ) Which one?_____

If Spanish is not your native language, how long have you been learning/speaking Spanish? ________ years ________ months

This survey comprises two sections:

Section I  Cultural Questionnaire

Section II  La posada

Please make sure you complete the TWO sections

Thank you very much
Appendix C

Cultural questionnaire translated to English
APPENDIX C
Section I
Cultural Questionnaire

Look at the questions below and circle the ONE number that is most appropriate for you. Please do not leave out any item. Thank you very much.

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at all</th>
<th>Not so much</th>
<th>So - so</th>
<th>A little</th>
<th>Quite a lot</th>
<th>Very much</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.- Do you find learning Spanish interesting?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2.- Do you find speaking Spanish interesting?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>3.- Do you enjoy learning Spanish?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>4.- Do you enjoy speaking Spanish?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>5.- Do you think Spanish is important in the world these days?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6.- Do you like the people who live in Latin America and Spain?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7.- Do you like speaking Spanish with a native Spanish-speaker?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8.- Do you think it is important meeting people from Spanish-speaking countries?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>9.- Do you like to travel to Spanish-speaking countries?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>10.- Would you like to know more about people from Spanish-speaking countries?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>11.- Are you interested in the values and customs of Spanish-speaking cultures?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>12.- Do you want to learn more about Spanish culture?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>13.- Do you like the music of Spanish-speaking countries?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>14.- Do you like Spanish films?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>15.- Do you like TV programmes made in Spanish-speaking countries?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
Appendix D

Head-verbs frequency rankings
## APPENDIX D

Head-verbs frequency rankings

<table>
<thead>
<tr>
<th>Mark Davies Verbs Rank Frequency List</th>
<th>Chang-Rodriguez Rank</th>
<th>Graded Spanish Word Book</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verb</strong></td>
<td><strong>Rank</strong></td>
<td><strong>Verb</strong></td>
</tr>
<tr>
<td>Tener</td>
<td>18</td>
<td>Tener</td>
</tr>
<tr>
<td>Dar</td>
<td>39</td>
<td>Dar</td>
</tr>
<tr>
<td>Volver</td>
<td>112</td>
<td>Volver</td>
</tr>
<tr>
<td>Contar</td>
<td>155</td>
<td>Contar</td>
</tr>
<tr>
<td>Perder</td>
<td>190</td>
<td>Perder</td>
</tr>
<tr>
<td>Faltar</td>
<td>510</td>
<td>Faltar</td>
</tr>
<tr>
<td>Echar</td>
<td>455</td>
<td>Echar</td>
</tr>
<tr>
<td>Sentar</td>
<td>710</td>
<td>Sentar</td>
</tr>
<tr>
<td>Dirigir</td>
<td>331</td>
<td>Dirigir</td>
</tr>
<tr>
<td>Entregar</td>
<td>558</td>
<td>Entregar</td>
</tr>
<tr>
<td>Quemar</td>
<td>1,509</td>
<td>Quemar</td>
</tr>
<tr>
<td>Besar</td>
<td>3,696</td>
<td>Besar</td>
</tr>
<tr>
<td>Lavar</td>
<td>1,762</td>
<td>Lavar</td>
</tr>
<tr>
<td>Ahogar</td>
<td>2,821</td>
<td>Ahogar</td>
</tr>
<tr>
<td>Consultar</td>
<td>1,587</td>
<td>Consultar</td>
</tr>
<tr>
<td>Aguar</td>
<td>*10,000</td>
<td>Aguar</td>
</tr>
<tr>
<td>Ahuecar</td>
<td>*10,000</td>
<td>Ahuecar</td>
</tr>
<tr>
<td>Empinar</td>
<td>*10,000</td>
<td>Empinar</td>
</tr>
<tr>
<td>Colar</td>
<td>4,237</td>
<td>Colar</td>
</tr>
<tr>
<td>Freir</td>
<td>*10,000</td>
<td>Freir</td>
</tr>
</tbody>
</table>

*Not appearing in any list
Appendix E

Frequency criterion
## APPENDIX E
Frequency Criterion

<table>
<thead>
<tr>
<th>Head Verbs</th>
<th>Category</th>
<th>Frequency Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL Tener</td>
<td>High Frequency Light Verbs</td>
<td>Appearing in the top 1-200 words</td>
</tr>
<tr>
<td>HL Dar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL Volver</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL Contar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL Perder</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H Faltar</td>
<td>High Frequency Lexical Verbs</td>
<td>Appearing in the 201-1000 words</td>
</tr>
<tr>
<td>H Echar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H Sentar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H Dirigir</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H Entregarse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M Quemarse</td>
<td>Medium Frequency Lexical Verbs</td>
<td>Appearing in the 1001-4000 words</td>
</tr>
<tr>
<td>M Besar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M Lavar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M Ahogarse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M Consultar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L Aguar</td>
<td>Low Frequency Lexical Verbs</td>
<td>Appearing after 4001 words or not</td>
</tr>
<tr>
<td>L Ahuecar</td>
<td></td>
<td>appearing in any list</td>
</tr>
<tr>
<td>L Empinar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L Colarse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L Freir</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix F

Final list of formulaic expressions
# APPENDIX F
List of Selected Formulaic Expressions

<table>
<thead>
<tr>
<th>Code</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL</td>
<td>Tener ángel</td>
</tr>
<tr>
<td>HL</td>
<td>Dar el primer paso</td>
</tr>
<tr>
<td>HL</td>
<td>Volver la vista atrás</td>
</tr>
<tr>
<td>HL</td>
<td>Contar el cuento</td>
</tr>
<tr>
<td>HL</td>
<td>Perder el tiempo</td>
</tr>
<tr>
<td>H</td>
<td>Faltar a su palabra</td>
</tr>
<tr>
<td>H</td>
<td>Echar flores</td>
</tr>
<tr>
<td>H</td>
<td>Sentar cabeza</td>
</tr>
<tr>
<td>H</td>
<td>Dirigir la palabra</td>
</tr>
<tr>
<td>H</td>
<td>Entregarse en cuerpo y alma</td>
</tr>
<tr>
<td>M</td>
<td>Quemarse las pestanas</td>
</tr>
<tr>
<td>M</td>
<td>Besar el suelo que (n) pisa</td>
</tr>
<tr>
<td>M</td>
<td>Lavar el cerebro</td>
</tr>
<tr>
<td>M</td>
<td>Ahogarse en un vaso de agua</td>
</tr>
<tr>
<td>M</td>
<td>Consultar con la almohada</td>
</tr>
<tr>
<td>L</td>
<td>Aguar la fiesta</td>
</tr>
<tr>
<td>L</td>
<td>Ahuecar el ala</td>
</tr>
<tr>
<td>L</td>
<td>Empinar el codo</td>
</tr>
<tr>
<td>L</td>
<td>Colarse en la fiesta</td>
</tr>
<tr>
<td>L</td>
<td>Freir espárragos</td>
</tr>
</tbody>
</table>