Learning multiword expressions from corpora and dictionaries

Orsolya Vincze

Doctoral thesis

2015

Supervisor:

Dr. Margarita Alonso Ramos

Departamento de Galego-Portugués, Francés e Lingüística

UNIVERSIDADE DA CORUÑA

Orsolya Vincze

Learning multiword expressions from corpora and dictionaries

Universidade da Coruña

Departamento de Galego-Portugués, Francés e Lingüística

Supervisor:

Dr. Margarita Alonso Ramos

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Margarita Alonso Ramos. I am really thankful for her guidance, time and dedication, not only when it comes to research, but also in matters of bureaucracy and life. Clearly, without her perseverance in standing up to and supporting her students, I would have abandoned my studies before I even started. She has taught me that research is done with rigor and requires certain sacrifice, but is, at the same time, exciting and rewarding.

I also express my appreciation for the funding received through the FPU grant (AP2010-4334) of which I was recipient during the completion of the thesis. The research presented in this thesis was carried out in the framework of the research project "*Herramienta de ayuda a la redacción en español: fundamentos lingüísticos para el procesamiento de colocaciones*" supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2011-30219-C02-01. This project was undertaken by a research group at the University of A Coruña directed by Margarita Alonso Ramos and a research group at the Pompeu Fabra University under the direction of Leo Wanner. I am greatly indebted to all members of both teams for their valuable contributions at different stages of the project. I am especially thankful to the members of the DiCE group, Estela Mosqueira Suárez, Ana Orol González and Marcos García Salido, who I have worked most closely with, and whose contribution and companionship has been invaluable for the development of this thesis, as has also been Nancy Vázquez Veiga's enthusiasm and wisdom in all matters of academia.

Similarly, I express my gratitude to Cristobal Lozano and Amaya Mendicoexea, who compiled the CEDEL2 corpus, and kindly provided us access to the data. I would also like to thank those who opened up their classrooms and offered help with the collection of data necessary for the experimental studies, especially Begoña Sanromán Vilas at the University of Helsinki, Elena Sánchez Trigo at the University of Vigo, Ángel Francisco Martínez Fernández at the EOI A Coruña and Montserrat Muriano Rodríguez at the University of A Coruña. I am also indebted to Miguel Ramos Naveira, who offered great help on the technological front, especially with the DiCE interface and the usability logs.

I am thankful for the support received from Averil Coxhead, who welcomed me at the School of Language and Applied Linguistics Studies of Victoria University of Wellington. Working in such stimulating environment has contributed immensely to my dissertation. I would also like to thank my former professors back in Hungary, at the University of Szeged, who introduced me to the exciting and multifaceted field of linguistics, among them, especially, Anna Fenyvesi and Tibor Berta.

My gratitude also goes to all my friends. I am especially grateful to my PhD "inmates", both for the valuable discussions and the encouragement in the more difficult times; Iria, Rike, Cath, Luz, María, Luis and, of course, all inhabitants of VZ410, thank you. Thanks also to Tito for pointing out repeatedly that a dissertation "is never finished, only abandoned". I am immensely grateful to my friends not – as yet – completing dissertations, who, regardless, endured my complaints and provided encouragement. Arturo, Noelia, the two Marías, Fer Vet, Mauro, Fran, Bea, Paula, Yulia, Éva, Judit, and several others, you have always been there for me.

Lastly, I would like to thank my family for all their love and support. I am very grateful to my parents for granting me complete freedom in pursuing whatever goals I preferred. Above all, I would like to express my gratitude to Chan for his love, constant support and encouragement during the final stages of this dissertation.

Abstract

The purpose of the present thesis is to examine Spanish as a foreign language (SFL) learners' needs when it comes to enhancing their collocation competence and use, with a view to designing an online collocation learning tool aimed at learners of Spanish. Accordingly, the research presented here corresponds to the following three aims. Firstly, SFL learners' collocation use is explored through a learner corpus study carried out using material from the CEDEL2 corpus. SFL learners' collocation use is compared to that of native speakers, while learner collocation errors are also examined. Secondly, the thesis examines the design and functionalities of existing learning tools that can support collocation learning, such as collocation dictionaries and corpus-based tools. More specifically, it describes a usability experiment focusing on the interface of the *Diccionario de colocaciones del español*, as well as a study testing SFL learners' ability to autonomously correct collocation errors with the help of concordance data obtained from corpus. Thirdly, taking into account the findings of these studies, the design of an online collocation learning tool aimed at SFL learners is described.

Resumen

El propósito de la presente tesis es examinar las necesidades de los aprendices de español como lengua extranjera (ELE) en lo que respecta el desarrollo de su competencia y uso colocacional con el objetivo de diseñar una nueva herramienta didáctica dirigida a aprendices de español. Por consiguiente, la investigación que presentamos corresponde a los siguientes tres objetivos principales. En primer lugar, exploramos el uso colocacional de aprendices de ELE mediante un estudio de corpus de aprendices que se ha llevado a cabo utilizando datos del corpus CEDEL2. Comparamos el uso de colocaciones de aprendices al de hablantes nativos del español, y, al mismo tiempo, examinamos los errores colocacionales de aprendices. En segundo lugar, la tesis examina el diseño y las funcionalidades de herramientas didácticas existentes que pueden ser aprovechados en el aprendizaje de colocaciones como son los diccionarios de colocaciones y herramientas basadas en datos de corpus. Más específicamente, presentamos un experimento de usabilidad del *Diccionario de colocaciones del español*, así como un estudio que examina la destreza de aprendices de ELE en corregir errores colocacionales autónomamente con la ayuda de concordancias obtenidas de corpus. En tercer lugar, teniendo en cuenta los resultados de estos estudios, describimos el diseño de una herramienta en línea centrada en colocaciones y destinada a aprendices de ELE.

Resumo

O propósito da presente tese é examinar as necesidades dos aprendices de español como lingua estranxeira (ELE) no que respecta ao desenvolvemento da súa competencia e uso colocacional, co obxectivo de deseñar unha nova ferramenta didáctica dirixida a aprendices de español. Por conseguinte, a investigación que presentamos corresponde aos seguintes tres obxectivos principais. En primeiro lugar, exploramos o uso colocacional de aprendices de ELE mediante un estudo de corpus de aprendices que se levou a cabo utilizando datos do corpus CEDEL2. Comparamos o uso de colocacións de aprendices ao de falantes nativos de español e, asemade, analizamos os erros colocacionais de aprendices. En segundo lugar, a tese examina o deseño e as funcionalidades de ferramentas didácticas existentes que poden ser aproveitados para a aprendizaxe de colocacións, como son os dicionarios de colocacións e ferramentas baseadas en datos de corpus. Máis especificamente, presentamos un experimento de usabilidade do *Diccionario de colocaciones del español*, así como un estudo que examina a destreza de aprendices de ELE na corrección de erros colocacionais autonomamente coa axuda de concordancias obtidas de corpus. En terceiro lugar, tendo en conta os resultados destes estudos, describimos o deseño dunha ferramenta en liña centrada en colocacións e destinada a aprendices de ELE.

Table of contents

List of tables and figures

List of abbreviations

BBI	BBI dictionary of English word combinations
BNC	British National Corpus
CEDEL2	Corpus Escrito del Español L2
COBUILD	Collins COBUILD Advanced Dictionary
CREA	Corpus de referencia del español actual
DDL	Data-driven learning
DiCE	Diccionario de colocaciones del español
ECL	Explanatory Combinatorial Lexicology
EFL	English as a foreign language
L1	First language or native language
L2	Second language
LCD	Longman Collocations Dictionary and Thesaurus
LDOCE4	Longman Dictionary of Contemporary English 4th edition
LDOCE5	Longman Dictionary of Contemporary English 5th edition
LF	Lexical function
LU	Lexical unit
MCD	Macmillan Collocations Dictionary
MI	Mutual information
MLD	Monolingual learners' dictionary
NPMI	Normalized pointwise mutual information
OCD	Oxford Collocations Dictionary 2nd edition
OCD1	Oxford Collocations Dictionary 1st edition
PCIC	Plan curricular del Instituto Cervantes
Práctico	Diccionario combinatorio práctico del español contemporáneo
Redes	Redes. Diccionario combinatorio del español contemporáneo
SFL	Spanish as a foreign language
SkE	Sketch Engine
SkELL	Sketch Engine for Language Learners
SLA	Second language acquisition

Chapter 1. Introduction

1.1 Motivation and context

Over the last few decades, interest in vocabulary has increased considerably within the field of second language acquisition (SLA). Importantly, lexical knowledge has come to be conceived of as not only comprising single word items, but also including a significant amount of multiword units. This approach is guided by claims analogous to the one formulated by Pawley and Syder (1983, 193), who observe that in language "only a small proportion of the total set of grammatical sentences are […] readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', odd or 'foreignisms'." Consequently, it is argued that native-like language cannot be attained through using single lexical items as individual units, hence learners should be introduced to an increasing number of multiword items or word combinations (see Sinclair 1991). These expressions are often referred to by different – and not necessarily overlapping – terms such as *chunk*, *formulaic language*, *formula, multiword unit*, *phraseme,* etc. (see e.g. Wray 2002, 8–9).

Evidence from a number of studies on learner language also points at the importance of enhancing second language (L2) learners' knowledge of multiword expressions. In fact, various authors studying learner corpora have arrived at the conclusion that insufficient or deficient use of formulas is one of the major factors separating the language use of advanced learners from that of native speakers (Kjellmer 1991; Granger 1998; Howarth 1998a). A review of relevant literature allows for identifying several arguments for teaching formulas to L2 learners: formulaic language is believed to contribute not only to more native-like production but also to increased fluency, while it is also held to be crucial for the general language acquisition process by advocates of formula-based learning (Durrant 2008, 40–57). In an attempt to explain why L2 learners fail to acquire formulaic language, Wray (2002) suggests that adult L2 learners' more developed cognitive systems and certain situational factors setting apart first language acquisition from second language learning cause learners to focus their attention on single words rather than on word combinations. In a different approach, Durrant (2008, 185), among others, claims that this problematic aspect of learner language is to be put down to the lack of sufficient input due to less exposure to the target language, affecting especially the acquisition of low frequency combinations (see also Henriksen 2013, 40–42).

This thesis aims to contribute to the research concerning language learners' knowledge and use of multiword expressions, as well as to explore ways in which these can be better presented in learning tools, contributing to more efficient learning. More specifically, the type of multiword expression targeted is that of *collocation.* The term *collocation* is used here, following Hausmann (1979; 1989) and Mel'čuk (1998; 2012), to designate restricted binary word combinations, such as e.g. *take a walk, heavy smoker,* etc. Since this kind of expression is considered of major importance for language learners, it has become the subject of an increasing number of studies in the field of SLA, including learner corpus research (e.g. Granger 1998; Howarth 1998; Nesselhauf 2005; Durrant and Schmitt 2009), experimental studies of collocation competence (e.g. Gyllstad 2007; Moreno Jaén 2009; Siyanova and Schmitt 2008), as well as studies concerning the evaluation and development of lexicographical works and teaching material focusing on collocations (e.g. Alonso Ramos 2008; Komuro 2009; Lew and Radłowska 2010).

1.2 The main objectives of the thesis

The purpose of the present thesis is to examine Spanish as a foreign language (SFL) learners' needs when it comes to enhancing their collocation competence and use. This involves an exploration of SFL learners' collocation use, as well as that of the learning resources they currently have at their disposal. The aim of the corresponding studies presented is to produce empirical evidence that can be exploited in the design of a novel online collocation learning tool aimed at learners of Spanish. Accordingly, three main questions will be addressed:

1. What is SFL learners' collocation production like?

The first question concerns the nature of SFL learners' collocation use. Since this work involves adopting a needs-driven approach to collocation learning, the starting point will be necessarily constituted by a description of the collocation production of the target learner group. Importantly, existing studies dealing with language learners' collocation production have concentrated almost exclusively on learners of English. While, in the case of Spanish, although proposals for teaching collocations do exist (e.g. Álvarez Cavanillas 2008; Ferrando Aramo 2009; Higueras García 2006), these are merely founded on the assumption that this type of multiword expression is generally problematic, with no solid empirical basis with respect to the collocation use of actual learners.

2. *What kind of collocation resources do language learners have and how successfully are these used?*

The second question aims at exploring what collocation learning resources are available to language learners in general, and SFL learners in particular, what these resources are like, and to what extent learners are able to exploit them. It has been commonly emphasized both in the more general case of vocabulary learning (e.g. Nation 2001), and collocation learning in particular (e.g. Hill, Lewis, and Lewis 2000; Nesselhauf 2005; Woolard 2000), that students should be equipped with strategies allowing for autonomous learning. Two types of resources frequently recommended for autonomous use and to support collocation production are dictionaries and language corpora. In addition, a number of online corpus tools tailored to learners' needs as well as more sophisticated collocation learning tools are also available. While the number of available resources targeting Spanish collocations is considerably lower than those aimed at English as a foreign language (EFL) learners, studies targeting usability issues related to these tools are generally scarce.

3. What should an online collocation learning tool designed for SFL learners be like?

The third question to be addressed concerns the design of an online collocation learning tool aimed at learners of Spanish. As suggested above, the aim of this thesis is to approach the creation of a learning tool from a needs-driven perspective, i.e. relying on the findings from empirical studies. We will see that existing resources are often insufficient in that they do not exploit the full potential of the electronic medium, do not observe all potential learner needs when it comes to producing collocations, or they are not suited to learners' reference skills.

1.3 Thesis outline

The present thesis consists of five main chapters. Chapter 2 and Chapter 3 provide a general theoretical background and review existing studies, while each of the remaining chapters addresses one of the above presented research questions through describing original work. Chapter 2 provides a review of the notion of collocation as defined from the perspective of different theoretical approaches, and examines the ways in which collocations have been described and classified by different authors. This discussion is especially relevant to establish the concept of collocation underlying the empirical work presented in the thesis. Chapter 3 gives an overview of the broad domain of inquiry relevant to the relationship of multiword expressions and foreign language teaching and learning. The first half of the chapter discusses the role attributed to multiword expressions in the language learning process, and examines empirical studies concerning L2 learners' collocation competence and use, while the second half is dedicated to describing pedagogical proposals concerned with collocation teaching, and, especially, to examining existing learning resources potentially applicable for autonomous collocation learning, such as dictionaries, corpora and online tools. Chapter 4 deals with the collocation use of language learners in written language. It presents a learner corpus study which aims at describing lexical combinations produced by SFL learners and examining learner collocation errors. Chapter 5 describes two empirical studies. The first of these constitutes a usability experiment evaluating the user interface of a Spanish online collocation dictionary, DiCE (Alonso Ramos 2004), while the second aims to assess language learners' ability to autonomously identify and correct collocation errors with the help of corpus examples. Finally, Chapter 6 describes the design of a novel collocation learning tool, taking into account evidence relevant to the needs of the target user group, i.e. SFL learners.

Chapter 2. Collocations in linguistic description

2.1 Introduction

The present chapter introduces the phenomenon of collocation, which constitutes the focus of this thesis. It aims to clarify the concept of collocation first through providing an overview of the different descriptive frameworks in which it emerged, and, second, trough examining its key features as described within the *Explanatory Combinatorial Lexicology* (ECL, see e.g. Mel’čuk et al. 1984; Mel’čuk et al. 1988; Mel’čuk et al. 1992; Mel’čuk et al. 1999), constituting the theoretical framework for the definition of collocation adopted in this thesis, as well as in the work of authors representing different descriptive traditions. Following this, the classification of combinations according to their syntactic properties and semantic content are discussed. These two aspects of the description of collocations are highly relevant in the creation of collocation dictionaries, and, consequently, in the pedagogical context.

2.2 Defining collocation

The first use of the term *collocation* to refer to a linguistic phenomenon, similar to what is designated by it in present day approaches, is often attributed to Palmer (1933) (see e.g. Mitchell 1971, 35). Nevertheless, despite the long history of the term, no general consensus exists in the current literature as to the exact characteristics of the type of expression the term refers to. This is especially apparent in that most studies dealing with collocations start by providing a corresponding definition. The field of SLA is no exception to the rule; as noted by e.g. Nesselhauf (2005, 3), in studies on collocations in learner language, “the use of the term is often hazy”. One clear drawback of this situation is that, whenever one intends to provide an overview or compare the results of different studies, each set of data has to be handled with caution, measuring the consequences of the approach adopted by the given author(s).

As it was mentioned above, this thesis adopts the concept of collocation established within the ECL framework. Nevertheless, in order to gain a fuller understanding of the phenomenon, it is necessary to observe the broader context of the study of collocations. Despite the varying definitions, it is commonly accepted that within linguistic research collocations have been approached from two main angles (see e.g. Nesselhauf 2004; Alonso Ramos 1994-1995; Gyllstad 2007a). One of these, most often adopted by corpus linguists and computational linguists, studies collocations as an empirical concept, and assumes that they constitute a directly observable property of language (Evert 2009, 1218). Thus collocations are defined roughly as units of two or more words co-occurring in texts, often with a given frequency. The other approach, to which the ECL framework can be ascribed, is generally attributed to the fields of lexicography and language pedagogy, and describes collocations as lexically significant word combinations, which constitute a part of speakers' linguistic competence. Following Nesselhauf (2004), I will refer to these two traditions as the *frequency-based approach* and the *phraseological approach*, respectively.

Sections 2.2.1 and 2.2.2 provide an overview of the concept of collocation in the frequency-based and the phraseological approaches. Section 2.2.3 discusses the definition of collocation and its relationship to other types of phraseological units defined within the ECL framework. Following this, Section 2.2.4 examines the key features of collocations comparing the ECL approach to descriptions provided by other authors, and Section 2.2.5 summarizes the characteristics of collocations following the definition used in the present thesis.

2.2.1 Collocations as frequent combinations of words

In the frequency-based approach, collocations are considered – in broad terms – as words co-occurring in a text at a given distance and with a certain frequency. This point of view is also often referred to as the *Neo-Firthian* tradition, since it is largely founded on Firths' (1957; 1968a; 1968b) pioneering ideas, followed most notably by Halliday (1966) and Sinclair (1966; 1987; 1991). In what follows, I discuss the work of these three authors, which, as we will see, each reflect somewhat differing conceptions of collocation, as well as an evolution in the use of the term, while I also make mention of more recent tendencies relevant in the frequency-based definition of collocations.

Firth is often referred to as the father of the term *collocation.* Although, historically it was not him who introduced it in linguistics, his ideas inspired the work of many scholars, thus leading to collocation becoming a well-established subject of linguistic inquiry. The frequently quoted claim "You shall know a word by the company it keeps!" (Firth 1968a, 179) is often used to summarize Firth's interest in collocation, which was motivated by the study of meaning in language. He proposed to study the meaning on different levels, one of which is ‘meaning by collocation’, which he explains by claiming: “One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*”. (Firth 1957a, 196) In Firth (1957 – originally written around 1951) he demonstrates the interest of the study of collocations in stylistics, while in Firth (1968b – written in about 1951-1952) he shows how collocations can be exploited for lexicographical purposes, providing a division of senses for the English verb *get* based on “the types of situation in which the collocations as wholes may be used” (p. 22.).

However, as noted by Nesselhauf (2005, 2), one does not find in Firth’s work a clear definition of what he understands by the term collocation, his description of the nature of these expressions being rather vague or imprecise, as well as changing throughout his writing. In general terms, Firth’s conception of collocation can be summarized as the co-occurrence of words with certain proximity in a text. Descriptions referring to more specific characteristics, such as the frequency of co-occurrence of elements, the number of words comprising a collocation and the nature of ‘word’ as a member of a collocation are often contradictory or variable.

As for the frequency or probability of co-occurrence, sometimes Firth appears to consider co-occurrences of words as collocations independently of whether they are frequent or not, classifying them as ‘general’ or ‘usual’ and ‘technical’ or ‘personal’ collocations (Firth 1957a, 195). At other times he seems to concentrate rather on habitual combinations, claiming that collocating words are ‘mutually expectant’ of each other: “Collocations of a given word are statements of the habitual or customary places of that word in collocational order” (Firth 1968a, 181). Concerning the number of words making up a collocation, the examples analyzed by Firth tend to be combinations of two or more consecutive words, sometimes even whole sentences, such as in the commonly cited examples *you silly ass, he is a silly ass, don’t be such an ass* (Firth 1968a, 179). It is true however, that his other emblematic example of *dark* collocating with *night* (Firth 1957a, 196) seems to be more in line with current conceptions of collocation as a combination of two lexical items. Finally, when it comes to interpreting what Firth understands by ‘word’ as a member of a collocation, at one point he claims that: "It is important, however to regard each word separately at first, and not as a member of a paradigm. The collocations of *light* (n.s.) separate it from *lights* (n.s.) and *light* (n.adj.) from *lighter* and *lightest.*" (Firth 1968a, 181). From this statement, as well as other examples given elsewhere in

Firth's writing, it can be deduced that he assumes collocation to be of co-occurring word forms. As we will see, scholars inspired by Firth's work elaborated considerably on these different points concerning the characteristics of collocation – contributing to the operationalization of the term and preparing the ground for formal collocation queries in current corpus studies.

While Firth's interest for collocations was motivated by the study of meaning, for Halliday (1966, 148–149), the main reason for analyzing co-occurrence patterns was to contribute to linguistic description, which – as he wrote – had attained such detail that it began to surpass the boundaries of grammar. His definition of collocation is as follows: "Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c" (Halliday 1961, 276).

Halliday's definition conveys considerably more precision than his predecessor's writings, as he elaborates explicitly on aspects concerning the nature of composing elements and probability of co-occurrence. While Firth conceived of collocations as co-occurring word forms, in Halliday's terms, collocations correspond to more abstract combinations of lexical items, such that e.g. *strong, strongly, strength* and *strengthened* are all regarded as the same item, and, consequently, *strong argument, he argued strongly, the strength of his argument* and *his argument was strengthened* are all considered instances of the same collocation (Halliday 1966, 151). Halliday (1966, 156) also provided an explicit definition of probability of co-occurrence, claiming that lexical analysis should focus on lexical restrictions, that is, "the extent to which an item is specified by its collocational environment", which can be accounted for by "the frequency of the item in a stated environment relative to its total frequency of occurrence". In other words, in order to establish the probability of co-occurrence of item *a* (the *node*) with item *b* (the *collocate*) one has to consider both the instances of co-occurrence and instances of individual occurrences. Note that, although, with referring to "n removes", Halliday's definition contains a formalized notion of proximity of the collocating lexical items (*span*), he does not elaborate on what the exact measure should be (see Halliday 1966, 152.); we will see that this aspect is dealt with by Sinclair in more detail.

Sinclair's view of language emphasizes the essential nature of the phenomenon of collocation, through relating it to what he claims to be the dominating principle in the production and interpretation of texts, the *idiom principle,* as opposed to what he refers to as the *open choice principle*. According to the idiom principle, “a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments” (Sinclair 1991, 110). Given Sinclair’s interest in corpus linguistics, i.e. analyzing large bodies of actual text, his work on collocations is characterized by an attempt for further operationalizing the concept, resolving practical problems reminiscent in his predecessors’ work.

Sinclair defines collocation as “the occurrence of two or more words within a short space of each other in a text” (1991, 170), diverging from Halliday in that, instead of lexical items, he conceives of a collocational relationship existing between ‘words’, by which he in fact seems to understand lemmas[1]. Similarly to both Firth and Halliday, Sinclair also considers co-occurrences as collocations regardless of their frequency, distinguishing between ‘unexpected’ or ‘casual’ and ‘frequent’ or ‘significant’ co-occurrences. He claims, nevertheless, that linguistic discussions are often limited to collocations showing statistical significance. Criteria for distinguishing significant collocations is outlined in Sinclair (1966, 418), where the author proposes a formula to calculate the probability of occurrence of a combination in a given span, taking into account the number of co-occurrences, the number of individual occurrences of the node, the span size and the total size of the corpus. As to span size, Johns and Sinclair (1974, 21–22) establish, on the basis of an analysis carried out on a large body of text, that about 95% of collocational influence of a node takes place at a distance of within four words to the left and to the right of the node, setting thus the optimal span size for the search of collocations at ±4.

As we have seen, Halliday and, especially, Sinclair advanced Firth’s initial and rather vague idea of collocation towards a much more operationalized concept, which can be studied using empirical methods. Thus, their work has laid the foundations for studying collocations through the use of computers allowing access to large bodies of corpora. Subsequently, the methodology used to identify collocations in corpus has evolved considerably, with current methods of measuring collocation strength involving the application of different statistical methods (see e.g. Manning & Schütze 1999; Evert 2005; Pecina 2009).

[1] “A lemma is what we normally mean by a ‘word’” (Sinclair 1991, 173).

However, as for the notion of collocation, similarly to what we find in the work of the early proponents of the frequency-based tradition, a certain amount of vagueness prevails in current studies, together with a notable lack of consensus on the exact characteristics of collocation. In this regard, Gries (2013, 39) observes that the way each author decides to select their own criteria is affected by the purpose of their work. He claims, furthermore, that it is imperative that researchers state clearly what they consider collocations to be, taking into account a set of six dimensions. These refer to 1) the nature of lexical elements that can constitute a collocation (e.g. words, parts of speech), 2) the number of elements that make up a collocation, 3) the criteria used to establish a frequency threshold for a combination to be considered to be a collocation, 4) the distance and (un)interruptability of the members of the collocation, 5) the lexical and syntactic flexibility of the elements of the collocation (e.g. word forms, lemmas, lexical items – in Halliday's sense), and finally, 6) the role of semantic non-compositionality and non-predictability in the definition (pp. 138-139).

As the last one of these criteria indicates, in current corpus research the definition of collocation is not necessarily limited to a purely frequency-based analysis. With the aim of restricting collocational analysis to combinations which prove to be of more interest to linguistic study, researchers working within the frequency approach tend to adapt criteria referring to the syntactic pattern and/or restrictedness of combinations. The condition that collocations have to constitute grammatically well-formed combinations or combinations corresponding to previously established syntactic patterns can serve to filter "non-sense" word strings from the rank of frequent combinations (see e.g. Kjellmer 1994, xiv–xv). For instance, using this criterion the sequences *day he*, *felt very* or *run but* in the sentence shown in (1), do not qualify as collocations, even though they are formed by adjacent words. As for limiting collocational analysis to restricted or unpredictable combinations, Herbst (1996, 383–384) claims that statistically significant collocations may not necessarily be lexically interesting. According to this author, although the combination *dark night* used by Firth to exemplify the phenomenon of collocation may be frequent and statistically salient, it demonstrates nothing more than a fact of the world, which is that nights are dark.

(1) Having worked during all day he wanted to go for a run but felt very tired.

As we will see, the issues of syntactic relatedness and lexical salience, as opposed to statistical salience, are also addressed by authors working within the phraseological approach. The research framework that adopts a concept of collocation which incorporates frequency as well as the notions grammatical well-formedness and restrictedness is often referred to as a *mixed* or *combined approach*; see e.g. Gyllstad (2007, 15–17) and Moreno Jaén (2009, 44–47). Some of the authors who themselves subscribe to the Firthian tradition, but whose research incorporates notions related to the phraseological tradition, include Greenbaum (1970; 1974), Kjellmer (1994) and Mitchell (1971), while more recent corpus-based collocation research within this framework involves, for instance, the work of Almela (2011), Kilgarriff and Tugwell (2001) and Williams (1998). Recent corpus-informed collocation dictionaries such as the *Macmillan Collocations Dictionary* (Rundell 2010) also make use of frequency, as it is apparent in the definition of collocation provided in the introduction of this particular dictionary: “the property of language whereby two or more words seem to appear frequently in each other’s company” (p. vii).

On a final note, it should be added that, although the study of collocations in the frequency-based tradition has emerged merely as a textual approach, as we will see in 3.2, current theories of psycholinguistics and language acquisition see statistically significant word co-occurrences as evidence for the psychological association between words (see e.g. Conklin and Schmitt 2012; Ellis 2002; Hoey 2005; Wray 2002).

2.2.2 Collocations as restricted lexical combinations

While Firths’ ideas of ‘collocational analysis’ were developed into an empirical method of analyzing recurrent word combinations in language corpora, the phraseological approach treats collocations as idiosyncratic multiword expressions and aims at describing them in reference works. The origins of this tradition are traced back to the work of H. E. Palmer and A. S. Hornby in the early 1930s, as well as to that of Russian phraseologists, particularly V. V. Vinogradov and N. N. Amosova (Cowie 1998a). A common thread in the ideas of these authors, and their followers, is their interest in language teaching and learner dictionaries.

Within the phraseological tradition, collocations are described as a type of *phraseological unit*, – other terms frequently used are *word combination, multiword unit, set expression, fixed expression, phraseme,* etc.[2] Different types of phraseological units are

[2] For a list of terms used by authors in different languages see Corpas Pastor (1996).

commonly conceived of as forming a continuum, ranging from completely transparent and commutable free word combinations to formally invariable, non-compositional idioms (see also Greenbaum 1974, 81; Cowie 1994, 3168). In this framework, collocations are generally defined as restricted but non-idiomatic combinations, thus they occupy the central part of this continuum, which, nevertheless, constitutes a rather fuzzy area, without clear-cut boundaries. This is why, authors working in the phraseological tradition, tend to describe the characteristics of collocations through contrasting them with other types of word combinations at the ends of this continuum: non-restricted free combinations, at one and, and idioms, at the other.

Note that, in the Spanish context, there is some debate concerning whether collocations belong to the domain of phraseology. For Bosque (2001, 21–25) and García-Page (2001, 89), for instance, collocations as non-idiomatic combinations should not be included in the realm of phraseology, limited to the description of idiomatic fixed expressions. On the contrary, Corpas Pastor (1996) defines phraseology as including non-idiomatic expressions, consequently, she describes collocations as a type of phraseological unit.

Given the great number of scholars who can be ascribed to the phraseological tradition, as well as the notorious lack of consensus concerning the terminology used to name the different types of expressions studied, this section is confined to exploring the origins of the tradition and to provide a brief overview of the work of the most prominent scholars making use of the term *collocation* within the phraseological framework. Thus, the following paragraphs review the contributions of Hausmann (1989; 1979; 1998), M. Benson, E. Benson and R. Ilson (1986b; 1986c; 1989; 1990), Cowie (1981; 1988; 1994), and Howarth (1996; 1998b; 1998a), together with the most relevant authors ascribed to the phraseological tradition in the context of Spanish linguistics, Alonso Ramos (1993; 1994-1995; 1998; 2006) Bosque (2001; 2004a; 2011), Corpas Pastor (1996; 2001) and Koike (2001; 2002). As it was mentioned above, Mel'čuk's approach to collocations, which was exhaustively applied to the description of Spanish collocations by Alonso Ramos, also belongs to the phraseological tradition. However, given that it constitutes the theoretical framework adopted in the present thesis, it will be discussed more thoroughly in a dedicated section (see 2.2.3).

Cowie (1998a) provides an overview of some of the most influential pioneering ideas in British and Russian phraseology from the 1930s to the 1960s. He claims that although in this period, the two groups of researchers were generally unaware of the work of their contemporaries, subsequently, Western scholars were considerably influenced by the ideas of their Eastern colleagues. The origins of the British tradition are traced back to Palmer's and Hornby's joint work on a pedagogically oriented research project initiated in the late 1920s, which had the aim of collecting and classifying multiword units. The project report, *Second Interim Report on English Collocations*, published in Japan in 1933, emphasized the widespread nature of collocations in language use, and the learning difficulty they represent.

In the *Interim Report* collocations were defined as combinations of two or more words which, as opposed to 'free phrases' or 'free combinations', have to be learned as a whole. Palmer's work is characterized by providing a detailed description of the syntactic pattern of the collocations studied; however, his and Hornby's approach to multiword expressions differed from more current ones since it did not account for the varying degrees of idiomaticity characterizing different groups of expressions, now considered to constitute different subtypes of the phraseological spectrum. (Cowie 1998a, 210–211). In other words, Palmer and Hornby's concept of collocation is different from the current use in that these authors did not distinguish between collocations and idioms. This, however, does not interfere with the fact that their work had an important effect in recognizing the importance of idiomatic expressions in foreign language pedagogy and laying the foundations for their description in learner dictionaries (Cowie 1998b, 8).

It was the work of a group of Russian scholars that addressed the issues of fixedness and semantic opacity, and, as a consequence, dealt with the grouping and description of expressions showing different degrees of idiomaticity, termed *phraseological units* for the first time, thus providing the foundations of modern phraseology. According to Cowie (1998a, 213–216), it was Vinogradov (1947) and Amosova (1969) whose ideas had more influence in the development of the concepts currently in use in phraseological research. Both of these scholars established categories to group different types of expressions, making use of the important distinction between *unmotivated* and *motivated* phraseological units. While the components of unmotivated expressions, such as e.g. *spill the beans*, are claimed to have no relation with the meaning of the whole combination, the meaning of motivated expressions can be deduced from the meaning of their elements. In particular, one type of motivated combination was described as containing a component used with its direct meaning, while the meaning of the other component is determined by the context. This special behavior can be observed in e.g. *small talk, small hours* and *small change*, where the sense of the adjective *small* varies depending on the noun. The description of this phenomenon, Vinogradov called *phraseologically bound* meaning, constitutes an important contribution to the field of phraseology (Cowie 1998a, 214–215).

From the late 1960s on, Western linguists, especially those with interest in language teaching and learner dictionaries, started to adopt the main ideas of Russian phraseology in describing and categorizing idiomatic word combinations. The following paragraphs summarize the ideas regarding the concept collocation formulated in the work of Hausmann, Benson et al., Cowie, and Howarth.

Hausmann (1979, 190–191) aims to delimit the concept of collocations in a way that it is suitable for lexicography and better suited to language learners' needs. He proposes to define collocation as a characteristic combination of two words corresponding to one of the six syntactic patterns listed in his paper *Le dictionnaire de collocations* (1989, 1010). According to Hausmann (1989), further defining features of collocations are restricted combinability, which allows to distinguish them from free combinations (e.g. *the book is useful*), and semantic transparence, which distinguishes them from idioms (e.g. *to pull sb's leg*). The author also emphasizes that, given these features, collocations constitute principally a production rather than a reception problem from the point of view of the language learner, and that they are not to be seen merely as a textual phenomenon, since they are part of the linguistic system.

An important contribution of Hausmann to the description of restricted combinations is his view – based on the Russian school of phraseology – that the elements of a collocation are found in an unequal relationship. One of them, he calls the *base* ('base'), is semantically autonomous, while the other element, the *collocate* or *collocator*[3] ('collocatif') is selected by the base, and receives its semantic identity only when part of the collocation. Thus, for instance, in a combination corresponding to the verb+noun pattern, such as e.g. *explode a myth*, the noun is the base and the verb is the collocate. Although this distinction in itself does not constitute a novel idea, Hausmann's' terminology has become widely used, and contributed to a better understanding of the nature of collocation.

[3] Although both of these are found in the literature as English translations of the original tern, the first one will be consistently applied in this thesis.

M. Benson, E. Benson and R. Ilson's concept of collocation was developed mainly in relation to their work on the *BBI dictionary of English word combinations* (Benson et al. 1986b). In this volume, collocations are described as recurrent, semi-fixed combinations, and are divided in two major groups, grammatical collocations and lexical collocations (p. ix), both corresponding to a set of predetermined syntactic patterns. Grammatical collocations are described as consisting of a dominant word being a noun, adjective, verb, etc. and a preposition or a grammatical construction, such that possible types include noun+preposition (*blockade against*), noun+to infinitive (*a pleasure to do*), and adjective+preposition (*angry at sb*). Lexical collocations are characterized as not containing any dominant element, given their components belong to the lexical categories of verb, noun, adjective and adverb. Corresponding syntactic patterns include e.g. verb+noun (*to override a veto*), adjective+noun (*a formidable change*) and verb+adverb (*affect deeply*) combinations. The classification of lexical collocations proposed by Benson et al. (1986b), as well as the status of grammatical collocations will be discussed in more detail in 2.3.1.

Cowie, whose interest towards word combinations was principally raised due to his work on phraseological learners' dictionaries (1981), is one of the most prominent researchers dealing extensively with collocations within the phraseological approach. He elaborated a typology of phraseological expressions, he refers to as *word combinations*, dividing them into two major groups: *formulae* and *composites*. The former group refers to expressions, such as *good morning* and *how are you*, whose meanings correspond to their discourse functions, while the latter category includes idiomatic word combinations with referential or propositional meaning, functioning below the sentence level, such as *spill the beans*, *dry run* and *do a U-turn* (Cowie 1988, 134–135). Composites are further subdivided into the categories of pure *idioms*, *figurative idioms*, *restricted collocations* and *free or open collocations*. These are perceived as forming a continuum from "transparent, freely recombinable collocations at one end to formally invariable, unmotivated idioms at the other" (1994, 3168).

Cowie defines the categories described along the continuum formed by composites through contrasting their characteristics with respect to two main criteria: specialization of meaning and restricted collocability. While idioms as a whole are generally immutable and semantically opaque, in the case of collocations, only one element has a more or less specialized or figurative sense, the other being used with its literal meaning (Cowie 1981, 227–228). The difference between free collocations and restricted collocations lies in that the elements of the first one are openly collocable in relation to each other, while the elements of restricted collocations have a more limited collocational range. To exemplify this, Cowie compares the free collocation *explode a bomb,* where the given meaning of the verb can be combined with any 'explodable' noun, with the restricted collocation *explode a myth*, in the case of which the specialized figurative sense of the verb is only available with a limited number of nouns, such as *myth, belief, idea, notion,* and *theory* (Cowie 1981, 226). A further important feature of collocations, according to Cowie (1994, 3169), is constituted by arbitrary limitations of choice existing between the elements of the combination, such as in the case of *cut one's throat*, *slash one's wrist*, **slash one's throat* and *cut one's wrist*. Finally, it should be noted, that similarly to Halliday (1966, see above), Cowie defines collocations as 'abstract composites' comprised of one or more lexemes or roots, such that the same combination may appear in different syntactic configurations, such as *strong argument*, *argue strongly* and the *strength of his argument*. (Cowie 1981, 230–231; 1994, 3169).

Howarth's treatment of phraseological units is heavily based on the categories established by Cowie, besides, – similarly to the rest of the authors mentioned so far – his work is also principally motivated by his interest in language pedagogy. His work places particular emphases on the pedagogical significance of collocations – until then largely neglected in the language-teaching context (1998b, 42). As for his typology of word combinations, Howarth distinguishes between the two main categories of *functional expressions* and *composite units*. Similarly to Cowie's classification, the category of composite units is further divided into *idioms* and *collocations*, which are in turn split into further subcategories: *free collocations*, *restricted collocations*, *figurative idioms*, and *pure idioms.* The exact definitions offered by Howarth (1996, 47) for each subtype of composite units, together with corresponding examples are shown Table 1.

Howarth (1996, 34) proposes to make use of six criteria in order to identify and distinguish between the four subtypes of composite units – which, similarly to Cowie, he envisions as ranging along a continuum. When classifying specific combinations, these six criteria are to be applied in the following order: 1) grammatical well-formedness, 2) institutionalization, 3) semantic transparency, 4) commutability, 5) semantic unity and 6) motivation. Note that the first two criteria demonstrate that Howarth's aim is not only to distinguish between different types of phraseological expressions, but also to devise a methodology allowing to identify them in corpora.

Category	Definition	Examples
Free collocations	Combinations of two or more words in which the elements are used in their literal sense.	*blow a trumpet, explode a mine, eat cheese*
Restricted collocations	Combinations in which one component is used in its literal meaning, while the other is used in a specialized sense. The specialized meaning of one element can be figurative, delexical or in some way technical and is an important determinant of limited collocability at the other. These combinations are, however, fully motivated.	*blow a fuse, explode an idea, shrug one's shoulder, make a decision*
Figurative idioms	Combinations which have figurative meanings in terms of the whole. They may permit arbitrary synonymous substitution of one or more elements. They have a current literal interpretation and are clearly motivated.	*blow your own trumpet, act a part 'pretend to be sb else', closed ranks*
Pure idioms	Combinations that have a unitary meaning which cannot be derived from the meanings of the components. They permit almost no substitution, and are unmotivated.	*blow the gaff, blow to kingdom come, by far and away*

Table 1: Types of word combinations as defined by Howarth (1996, 47)

Much the same way as we have seen in the case of Cowie, the third and fourth criteria, semantic transparency and commutability are used to differentiate between free combinations and restricted collocations on the one hand, and restricted collocations and idioms on the other. According to Howarth, free combinations have a transparent meaning, which is derivable from the meanings of their elements and the syntactic pattern of the expression. While, unlike idioms, restricted combinations do not constitute single semantic units, at least one of their elements is used in a specialized sense. When it comes to commutability, the co-occurrence of the elements of a free combination can be predicted from their individual meanings, and their substitution generally has only logical limits, such that e.g. the verb *explode* has to be combined with a noun that refers to something 'explodable'. However, there might be arbitrary limitations affecting the combinability of free combinations. For instance, the idiom *kick the bucket* can only be used with a human subject despite the fact that its meaning does not express 'die in a human way'. In comparison, restricted collocations range from unalterable idiom-like expressions, such as *curry favor,* to combinations where the specialized sense of one element is determined by the other, and either one or both elements are commutable, e.g. *table a motion/a bill/an amendment* or *carry out/conduct an experiment/a test/a survey.*

Despite the strong interest demonstrated by the work of the above mentioned scholars in the field of phraseology and collocations in particular, in the context of the study of Spanish language, the term *collocation* had not been adapted until much later (Alonso Ramos 2002, 67). It was Alonso Ramos (1993; 1998), who first dedicated attention to the extensive study of the phenomenon of restricted lexical co-occurrence in Spanish, making use of the term. Subsequently, collocations and restricted combinations have become the subject of an increasing body of research. The most notable descriptive studies include the work of Alonso Ramos (1993, 1998, 2006), Corpas Pastor (1996, 2001) and Koike (2001), as well as the combinatorial dictionaries coordinated by Bosque and further studies published by the same author (2001, 2004a, 2004b, 2011).

Alonso Ramos' (1993; 1998) work is carried out within the framework of the ECL, discussed in the following section. When adopting the notion of collocation formulated within the ECL framework, the author emphasizes that these expressions should in all cases be identified and defined from the point of view of production, considering the potential difficulties they can cause for a non-native speaker. While collocations can be more or less transparent, constitute semantically restricted combinations or not, what they have in common is that they express meanings commonly associated with the base – often described by lexical functions (see 2.3.3) –, and may result unpredictable from the point of view of a language learner (see e.g. Alonso Ramos 2006).

In Corpas Pastor's (1996, 66) monograph on phraseology, collocation is defined as a phraseological unit, constituted by two lexical units in syntactic relationship, which by itself does not constitute a speech act or an utterance, and is characterized by combinatory restrictions given its fixed nature in what refers to the norm. The first part of this definition distinguishes collocations from what the author calls *phraseological utterances* (*enunciado fraseológico*), such as slogans, proverbs or speech acts. The second part makes reference to Coseriu's (1962) distinction between *system* and *norm*, in that, according to Corpas Pastor (1996, 51), while idioms can be described as fixed expressions with respect to the system, the fixedness of collocations is to be interpreted with respect to the norm.

In a subsequent study, Corpas Pastor (2001, 43–46) describes the five main traits of collocations as follows. She claims that 1) collocations are multiword expressions composed by two elements, one of which itself can be a multiword expression, i.e. an idiom (*llorar a moco tendido* 'to weep uncontrollably'); 2) their components co-occur with certain frequency; 3) they are institutionalized or fixed expressions, which are reproduced in discourse, and constitute a “psychological correlate” (see Greenbaum 1974, 83) for the speakers; 4) the institutionalization of a collocation translates into restricted combinatory, which often implies the semantic specialization of their components; and finally, 5) collocations are variable with the existence of completely synonymous combinations, such as *pegarse un susto, darse un susto* and *llevarse un susto* ‘to get scared’, stylistic variants, as *hacer una pregunta* (informal) and *formular una pregunta* (formal) ‘to ask a question’, and dialectal variants, e.g. *dar una opinion* (Spain) and *entregar una opinión* (Chile) ‘to give an opinion’.

Koike’s (2001) monograph dealing with the description and classification of Spanish collocations offers six criteria which closely resemble the concept of collocation described by the previous author. According to Koike, collocations can be defined as expressions characterized by 1) frequent co-occurrence of lexical units, 2) combinatory restrictions, 3) compositionality, 4) having a link between the components, 5) typicality of the relation between the components, and 6) semantic accuracy of the combination. The last two of these features deserve some explanation. Inspired by the concept of *lexical solidarities* introduced by Coseriu (1977), Koike aims to determine collocations in terms of the semantic nature of the relationship existing between their components, claiming that elements of a collocation represent a typical relationship in the “real world”. This means that e.g. *cargar una pistola* ‘to load a gun’ is a collocation, whereas *olvidar una pistola* ‘forget a pistol’ is not, given that the latter refers to an action not typically related to a shot gun. Another important semantic feature described by Koike lies in that collocations express a very definite context in a concise and precise way, such that e.g. the meaning ‘make somebody remember something they had forgotten’ is expressed by the combination *refrescar la memoria* ‘to refresh one’s memory’. Recall that, as mentioned earlier, collocations expressing typical meanings related to the core element (base) is the idea behind Mel’čuk’s lexical functions (see Alonso Ramos 2006, 40).

Note that, since they focus on the description of collocations as phraseological units, Corpas Pastor’s and Koike’s work is discussed within the phraseological approach, nevertheless, they also both postulate that collocations constitute frequent combinations. It has already been discussed in 2.2.1 that, while criteria used in the phraseological tradition, such as syntactic relationship between the components or restricted combinability are often used in the frequency-based approach to narrow the scope of collocational analysis to combinations presenting more linguistic interest, frequency information is on occasions applied as a way to operationalize the concept of collocation within the phraseological approach, such as in the case of the compilation of collocations dictionaries.

The last prominent author focusing on Spanish word combinations to be discussed here is Bosque, whose notion of collocation diverges in several aspects from that of the authors mentioned above. As opposed to Corpas Pastor and Koike, and in accordance with other authors working within the phraseological framework[4], Bosque (2001) maintains that frequency does not necessarily provide linguistically interesting data. Consequently, he claims that collocations should be defined as cases of lexical restriction, in his own terminology *lexical selection*, instead of frequent combinations. However, he diverges from e.g. Hausmann in perceiving the directionality of lexical selection, claiming that lexical restriction is imposed by the *collocate* or, following the author's terminology, the *predicate* on the *base* or *argument*. In addition, lexical selection as defined by Bosque, instead of being arbitrary, is of semantic nature, so that the collocate not only selects for a stand-alone base, but for groups of bases constituting a *lexical class,* which can be described in terms of semantic features. This means that e.g. the verb *prestar* 'lend, give' is combined with a number of nouns which can be characterized as having related meanings, such as *ayuda* 'help', *apoyo* 'aid', *atención* 'care', *asistencia* 'assistance', *colaboración* 'collaboration', and *servicio* 'service' (Bosque 2001, 15–20; 2004a, XCV–XCVI). Bosque's notion of restricted lexical combination and lexical selection is closely intertwined with the combinatorial dictionary *Redes. Diccionario combinatorio del español contemporáneo* (*Redes*, Bosque 2004b), and is further discussed in 2.3.2.1 and in 3.4.2.1.

2.2.3 Collocations in the Explanatory Combinatorial Lexicology

The notion of collocation adopted in the present thesis is the one formulated within the framework of the *Explanatory Combinatorial Lexicology* (ECL), which constitutes the lexical component of Igor Mel'čuk's language model represented by the Meaning⇔Text Theory (see e.g. Mel'čuk and Žolkovskij 1970; Mel'čuk 1981). As it was mentioned above, this theoretical framework can be placed within the phraseological tradition, described in the previous section. Here I present the definition of collocation, as well as the full typology of phraseological expressions offered within the ECL.

[4] Note that Bosque (2001, 10) in fact argues against including collocations in the field of phraseology, which for him is limited to the study of idiomatic expressions.

The Meaning⇔Text Theory proposed by Mel’čuk aims at providing a description of language through a multilevel linguistic model and a system of rules describing the correspondences between each consecutive level, starting from linguistic meaning to actual linguistic expressions, that is, text. An important feature of this model is that it attributes a great importance to lexis, providing highly formal instruments for its representation in what is called *Explanatory Combinatorial Dictionary* (ECD, see e.g. Mel’čuk et al. 1984; Mel’čuk et al. 1988; Mel’čuk et al. 1992; Mel’čuk et al. 1999). This framework puts particular emphasis on the description and classification of *phrasemes*, the term used by Mel’čuk to refer to the generic category encompassing all multiword expressions that constitute the object of study of the field of phraseology. Both the definition of phraseme and the classification of its subtypes have evolved from earlier works to more recent publications of the author; the overview provided here is limited to consider only the more current accounts.

Mel’čuk (2008; 2012; 2013) conceives of phrasemes as linguistically constrained multiword phrases which can be delimited through their contrast with *free phrases*. Thus phraseme is defined as a phrase, at least one of whose lexical components is selected in a linguistically constrained way (2012, 33), or simply as a non-free multilexemic expression (2013, 131). ‘Linguistically constrained’ essentially refers to the fact that the selection of one or more lexical component of the expression is determined or limited by another component of the same expression, consequently, the lexical item in question cannot be freely replaced by any of its synonyms without affecting the meaning and the grammaticality of the expression.

Collocations constitute one subtype of the category of phrasemes. The definition of collocation provided by Mel’čuk (1998, 29) is as follows:

> A COLLOCATION **AB** of language **L** is a semantic phraseme[5] of **L** such that its signified 'X' is constructed out of the signified of one of its two constituent lexemes -- say, of **A** -- and a signified 'C' ['X' = 'A⊕C'] such that the lexeme **B** expresses 'C' only contingent on **A.**

This definition summarizes in a concise and formalized manner Mel’čuk’s understanding of the notion of collocation. First of all, collocations are described as lexically restricted and compositional multiword units. These two criteria, i.e.

[5] Note that the term *semantic phraseme*, as it will be discussed briefly, has become obsolete in the most recent version of the typology of phrasemes proposed by the author Mel’čuk.

restrictedness and compositionality, are used in general to distinguish between the main types of phrasemes: *clichés, collocations* and *idioms*. Secondly, collocations are further characterized as consisting of two major elements, **A**, or, using Hausmann’s (see above) terminology, the *base*, freely chosen by the speaker, and **B**, the *collocate,* whose choice is lexically determined by the base to express a given meaning bearing on it. Thus, for instance, in the collocations *heavy* <u>*accent*</u> and <u>*armed*</u> *to the teeth,* the underlined elements constitute the base, while the adjective *heavy* and the phrase *to the teeth,* both expressing a meaning of intensification, function as collocates. Thirdly, the definition states that the meaning expressed by the collocate within a given combination does not necessarily coincide with the meaning it has outside a collocation. This observation echoes Vinogradov’s original idea concerning phraseologically bound meaning, also taken up by Cowie and Howarth (see above). As it is discussed in more detail in 2.2.4.6, Mel’čuk (1998, 29–30) classifies collocations in four major categories according to the relationship between the meaning of the collocate (‘B’) and its meaning in the given combination (‘C’).

In order to better understand Mel’čuk’s notion of collocation, it is worth considering how this type of multiword unit compares to other types of phrasemes distinguished by his typology. As suggested above, the author uses two main criteria to define different types of phrasemes: 1) *restrictedness* and 2) *compositionality*, allowing him to discriminate between the major types of phrasemes being a) *lexical-semantic phrasemes* or *clichés* and b) *lexical phrasemes*, which include *idioms* and *collocations*, see Table 2.

As it is discussed in more detail in 2.2.4.5 and 2.2.4.6, these two criteria are similar, but, at the same time, somewhat different from those used by other authors in the phraseological approach. Firstly, concerning restrictedness, Mel’čuk distinguishes between restrictions on the selection of lexical and semantic components, as well as on the situation of use of the expression. Secondly, in defining compositionality, he emphasizes that it should be interpreted as semantic restrictions applying to the combination of components (Mel’čuk 2012, 35–36; Mel’čuk 2013, 132–133), and not as equivalent to semantic transparence. In fact, Mel’čuk, highlights that opacity and transparence are subjective and gradable, contrarily to compositionality which is an objective and discrete feature (Mel’čuk 2013, 133).

				Type of constraint			**Compositionality**	**Examples**
				Lex.	**Sem.**	**Prag.**		
PHRASEMES	**LEXICAL-SEMANTIC**	***CLICHÉS***	Pragmatemes	+	+	+	+	*Hold the line!* (in a phone conversation); *emphasis mine* (after a quotation)
			Pragmatically non-constr. clichés	+	+	-	+	*A watched pot never boils; Red Planet* (‘Mars’); *The Moonlight Sonata*
	LEXICAL	***IDIOMS***	Full idioms	+	-	-	- Does not include the meaning of any of its lexical components	*by heart* ‘remembering verbatim’; *spill the beans* ‘reveal a secret’
			Semi-idioms	+	-	-	- Includes the meaning of one of its components, but not that of the rest, and it includes an additional meaning as semantic pivot[6]	*Private eye* ‘private detective’; *sea dog* ‘person with a lot of experience of ships and sailing’
			Quasi- or weak idioms	+	-	-	- Includes the meaning of both its lexical components and an additional meaning as its semantic pivot	*Start a family* ‘conceive a first child with one’s spouse, hence starting a family’
		COLLOCATIONS	Standard collocations	+	-	-	+	*Land a job; high winds; crack a joke, do a favor; strong coffee*
			Non-standard collocations	+	-	-	+	*Leap year; black coffee*

Table 2: Typology of phrasemes according to Mel’čuk (2012, 32–42; 2013, 132–145)

Lexical-semantic phrasemes or *clichés* are described as fully compositional expressions in the case of which both semantic and lexical restrictions apply. This means that clichés express a complex meaning which is obligatorily used in a given language, while other meanings and/or formulations are not admissible. As shown in Table 2, the group of clichés can be divided into further categories according to whether the expression is pragmatically constrained or not. Pragmatically constrained clichés or *pragmatemes* are thus also constrained by the situation of use. For instance, in English the conventional way to indicate on a sign that an object has been recently painted is *Wet paint*, whereas a sign stating $^{\#}$*Caution, painted* would be inappropriate, although it constitutes a grammatical phrase. The official statement we find on the container or packaging of perishable food

[6] Semantic pivot is defined in Mel’čuk (2012a, 36; 2013, 135) as the part ‘σ_1’ of meaning ‘σ’ (‘σ’ = ‘σ_1’ $\oplus$ ‘σ_2’) where ‘σ_2’ is a predicate and ‘σ_1’ is its argument. Importantly, the semantic pivot does not necessarily coincide with the lexical meaning of a component of a multiword expression. For example, in *take a shower,* the semantic pivot is ‘shower’, while in *sea dog* ‘person with a lot of experience of ships and sailing’ the semantic pivot is ‘person’.

products is another example of a pragmatically constrained cliché; in English it reads as *best before*, while in Spanish it is *consumir preferentemente antes de* 'consume preferably before' and in Hungarian *minőségét megőrzi* 'maintains its quality'. Other examples are: *in other words, to make a long story short,* etc. (2012, 33–36; 2013, 132–133).

As for the category of lexical phrasemes, which includes both collocations and idioms, as shown in Table 2, only lexical restrictions apply. This implies that, the meaning of these phrasemes is freely constructed by the speaker, whereas the lexical components to express the given meaning cannot be chosen freely, i.e. some or all are chosen depending on the rest. (2012, 33–34; 2013, 132)[7].

In the same way as with other authors working within the phraseological approach, collocations are distinguished from idioms on the basis of compositionality. Mel'čuk defines compositionality, as components of an expression being joined together according to the general rules of language, distinguishing between semantically compositional and non-compositional expressions. The combination *call sb's attention to sg,* for instance, is compositional, and thus is a collocation, since its semantic content is distributed in a natural manner among its components, i.e. *call* 'cause that X be concerned with', and *attention* 'attention' (Mel'čuk 2012, 35–36; Mel'čuk 2013, 132–133).

Collocations are further classified depending on the specific meaning contributed by the collocate to that of the combination. The category of *standard collocations* includes combinations in which the semantic relation holding between the base and the collocate is applicable to many other bases, and defines many other collocates in the same language. For instance, the general meaning of 'intensification' is applied to several bases and can be expressed by a great number of different collocates: *heavy rain, strong tea, armed to the tooth, deep sleep.* In contrast, in the case of *non-standard collocations*, the semantic relation holding between the collocate and the base is applicable only to one base or a few bases in the given language, and defines only one collocate or a few collocates. For instance the meaning 'having 366 days' expressed in the combination *leap year* applies only to the given base and collocate, and the meaning 'with no dairy product added' is expressed by *black* only in combination with the base *coffee.* Standard collocations are

[7] In earlier accounts, Mel'čuk (1998, 28–29; 2008, 189–190) makes a primary distinction between *pragmatic phrasemes* and *semantic phrasemes*, explaining that in the former restrictions apply between the conceptual and the semantic levels of representation – following the MTT language model –, while in the latter, constraints are observed between the semantic and the deep syntactic levels of representation.

formally described by *standard lexical functions*, while non-standard collocations are represented by *non-standard lexical functions*, see 2.3.3 (Mel'čuk 2012, 39–40; 2013, 139–140).

In contrast to collocations, idioms are non-compositional. For instance, in the case of *kick the bucket,* the meaning 'die' is expressed by the whole expression, and not by any of the components. Further subcategories of idioms are established according to their degree of transparence/opacity: *full idioms* (e.g. *by heart* 'remembering verbatim'), semi-idioms (e.g. *private eye* 'private detective'; *sea dog* 'person with a lot of experience of ships and sailing') and *quasi-* or *weak idioms* (e.g. *start a family* 'conceive the first child with one's spouse, thus starting a family'). Table 2 demonstrates the semantic properties of these three types of idioms in more detail.

As it is shown here, Mel'čuk's approach to phraseology is rather powerful in the sense that it distinguishes between the different types of phrasemes using well defined and discrete criteria, providing a both theoretically well-founded and operationalizable framework to the study of collocations. An important feature of Mel'čuk's work is that it accounts for the characteristics of phrasemes through carefully discriminating between semantic and lexical phenomena. Even though – as generally noted by other authors, such as e.g. Cowie, Howarth and Bosque – certain correlation can be observed between combinatorial restrictions applying in the case of a combination and its semantic properties – a more discrete examination of these allows for a better understanding of the true nature of multiword units. Mel'čuk's method allows to define clearer boundaries between the different types of phrasemes, and, consequently, to avoid relying on the notion of the phraseological continuum, as is done by other authors working in the phraseological approach,.

2.2.4 The main characteristics of collocations

The previous sections mentioned a number of features which are used to define and describe the characteristics of collocations within the frequency-based approach and the phraseological approach, including the ECL framework. The present section discusses these key features one by one, through comparing the points of view of different authors, and contrasting them with the notion of collocation as understood in the present thesis.

2.2.4.1 Frequency of co-occurrence

Within the ECL framework, frequency of co-occurrence has no role whatsoever in the definition of collocations, since these are defined and contrasted to other types of phraseological units solely on the basis of criteria referent to the nature of restrictions on the selection of the components and to the compositionality of meaning of the expression. This is also the case also of other authors whose work is representative of the phraseological approach. According to Cowie (1998a, 226), for instance, “there is no clear evidence to date of a close correlation between measured frequency of occurrence and collocational restriction”.

In contrast, as it was discussed in 2.2.1, within the frequency-based approach to the study of collocations, information regarding the frequency of co-occurrence of two or more lexical elements is essential in describing the characteristics of the given combination, and/or in determining whether it is a collocation or not. In fact, the interpretation of frequency of co-occurrence as a definitorial criterion has given rise to gradually more refined techniques measuring the significance of combinations, which are used for identifying collocations in language corpora (see Sinclair 1966, 170; Manning and Schütze 1999; Kilgarriff and Tugwell 2001; Evert 2005; Pecina 2009). Nevertheless, authors tend to emphasize that the use of frequency as a sole criterion to single out collocations does not render sufficiently clean data for linguistic analysis, lexicographical or pedagogical purposes (e.g. Hausmann 1979, 190–191; Alonso Ramos 1993, 147–148; Kjellmer 1994, xiv; Herbst 1996, 383–384; Howarth 1996, 29; Cowie 1998a, 226; Bosque 2001, 11–15).

Consequently, in recent studies, certain convergence of criteria coming from the frequency-based and phraseological approaches can be observed when it comes to defining and/or identifying collocations. In more recent corpus-informed studies, one generally finds that, while a frequency threshold is used to establish whether a given combination is to be considered to be a collocation or not, other criteria, such as restrictions according to syntactic pattern or restricted lexical selection or semantic criteria are used to filter data (e.g. Kjellmer 1994; Kilgarriff and Tugwell 2001; Moreno Jaén 2009). At the same time, corpus technology and frequency information is often used to operationalize the notion of collocation, to validate or empirically support data in phraseologically or lexicographically oriented studies (see Martelli 2006; Nesselhauf 2005; Vincze and Alonso Ramos 2013). Finally, it should be noted, that the potential relevance of frequency of co-occurrence is reinstated by a number of researchers in the fields of psycholinguistics and language acquisition, who claim the existence of a relationship between frequency and language processing and/or language acquisition, as it is discussed in more detail in 3.2 (see e.g. Conklin and Schmitt 2012; Ellis 2002; Hoey 2005; Wray 2002).

2.2.4.2 Number of components

According to Mel'čuk's (1998, 29) definition – similarly to that of Hausmann (1989, 1010) –, collocations are composed of strictly two lexical units. Definitions provided by other authors are less restrictive regarding the number of components. We have seen that Firth, whose notion of collocation can be characterized as rather vague, includes virtually all kinds of co-occurrences or sequences of consecutive words under his collocational analysis, c.f. *you silly ass, he is a silly ass, don't be a silly ass* (Firth 1968a, 179). A similar view is represented by Kjellmer (1994, xxxvi), who admits in his dictionary frequent sequences of any number of consecutive word forms. While other authors, including both those belonging to the frequency-based tradition and those whose work is interpreted within the phraseological framework, define collocations as combinations of "two or more words" (e.g. Aisenstadt 1981, 54; Sinclair 1991; Cowie 1994, 3169; Howarth 1996, 47; Nesselhauf 2005).

Much of this apparent difference can be explained, however, by the fact that some of the authors do not define the nature of components or specify their role in the same way as it is done by Mel'čuk. As mentioned above, according to this author, the phenomenon of restricted lexical selection is observed between two lexical units. For instance, the string *prestar atención a* 'to pay attention to' is interpreted as a collocation where the selection of the verb *prestar*, the collocate, is imposed by the noun *atención*, the base, in contrast with the case of other authors who describe this combination as containing three lexical elements (i.e. words or word forms). For Benson et al. (1986a), the above mentioned string constitutes a combination of a lexical collocation (*prestar atención*) and a grammatical collocation (*atención a*). Note that the restricted selection involving the noun and the preposition is not considered to represent a collocational phenomenon in the ECL given that the preposition does not introduce any semantic content, but belongs to the government pattern of the noun. In the same manner, the ECL approach also describes a collocation containing an article, such as *tomar el sol* 'to sunbathe, lit. take the sun' is as a combination of two elements, the noun restricting the selection of the verb. Note that this does not imply that the use of governed prepositions or the obligatory use of an article in a given combinations is not specified in the dictionary.

Another difference with other approaches is that Mel'čuk defines collocation as a combination of two lexical units (see below), instead of words or word forms. Consequently, in the ECL framework, combinations termed *complex* collocations by e.g. Koike (2001, 44–60) and Corpas Pastor (2001, 43), are described as restricted combinations containing two lexical components, independently of the number of word forms. For instance, the expression *seguir a pie de letra* lit. 'follow by the foot of the letter' 'to follow carefully', is described as a collocation, with the verb *seguir* restricting the selection of the collocate, in this case, an idiom functioning as an adverbial, *a pie de letra*.

2.2.4.3 Nature of components

According to Mel'čuk's definition, the components of collocations are *lexical units* (LU), i.e. single or multiword items corresponding to a lexical meaning. This coincides with the notion of collocation described by Hausmann (1979; 1989) and Benson et al. (1986c).

In contrast, we have seen that Firth (1968a) conceived of collocations as combinations of particular word forms. Although this idea is not generally followed in the recent literature, and especially not the phraseological approach, some authors have argued that it is indeed important to take into account the fact that certain combinations not necessarily admit all inflected forms of a lemma. Moreno Jaén (2007, 54–55), for instance, claims that there is an important difference between the two main elements of the collocation in this respect, given that the form of the base is often restricted, whereas that of the collocate is not. For instance, the noun *eye* in its plural and singular forms has distinct combinatorial characteristics, as illustrated in *the naked eye* and *to roll one´s eyes*, whereas it is of no major interest to consider forms such as *I made a decision* and *she makes a decision* as different collocations (see also Kjellmer 1994, xix). From the point of view of the ECL framework, restrictions applying to the form of composing lexical items do not justify taking the word form as the composing element of the collocation, nevertheless, they should be specified in the lexicographic description of the combination, together with its morphosyntactic properties, such as the use of articles, prepositions (see 3.4.2.1.C and 6.2.1.2).

Halliday (1966) represents yet another different approach through defining collocations as abstract lexical items, whereby e.g. *strong argument, to argue strongly, the strength of an argument* and *to strengthen an argument* are considered to be instances of the same combination (see Section 2.2.1). Following this idea, Mitchell (1971, 50) claims that collocations are “of roots, not of words”, such that *hard-working, to work hard, hard work* and *hard worker* represent the same combination. In the same vain, Cowie describes collocations as “abstract composites” comprised of one or more lexemes or roots (1994, 3169), thus, this author also maintains that the same combination may appear in different syntactic configurations, such as *strong argument, argue strongly* and *the strength of his argument* (1981, 230).

While this approach is admittedly useful in accounting for the productivity of certain combinations, generalizations, such as those represented by Halliday’s, Mitchell’s or Cowie’s examples, are not possible in all cases, as pointed out by Greenbaum (1974, 81). The latter author claims that the description of collocations that “disregards word-class categories” is problematic due to restrictions in the case of many combinations such as e.g. *badly need,*bad need* in comparison with *desperately need and desperate need*. Note that, the MTT framework, whose lexical model is constituted by the ECD, has a mechanism which enables stating “semantic equivalence between different lexico-syntactic constructions”. So called *paraphrasing rules* are formulated in terms of lexical functions (see 2.3.3), used to represent lexical relations, including collocations (see Iordanskaja et al. 1996, 289–291).

In sum, describing collocations in terms of lexical units, as it is suggested within the ECL framework, is more economical than considering combinations of word forms, at the same time, it leaves room for describing morphological restrictions, as well as establishing correspondences between lexically related combinations. Furthermore, as we will see in 2.3.3, lexical functions used to describe collocations allow to group synonymous expressions with parallel syntactic patterns together: *strong/powerful/sound argument, to argue strongly/passionately/forcefully*.

2.2.4.4 Relationship between the components

Definitions provided by the authors whose work on collocations has been reviewed so far characterize the relationship holding between the elements of a collocation in different ways. As mentioned above, the only condition originally set out by the Firthian approach was that elements had to co-occur, sometimes with certain proximity, in texts. In order to narrow down the list of combinations to be studied, other conditions, such as grammatical well-formedness and the existence of combinatorial restrictions, were introduced. Both of these have implications on how the relation between the components of a collocation is described. The condition of grammatical well-formedness implies that collocations have to correspond to a certain syntactic pattern, i.e. a certain grammatical or syntactic relationship must hold between the components (e.g. verb-object, noun-modifier, etc.), while the description of combinatorial restrictions led certain authors to claim the existence of directionality regarding which element imposes the selection of the other (see Greenbaum 1970; Greenbaum 1974; Kjellmer 1984; Kilgarriff and Tugwell 2001; Mitchell 1971; Williams 1998).

Although it is not explicitly stated in the formal definition cited in 2.2.3, Mel'čuk also expects the composing elements of the collocation to be bound by a syntactic relation. In fact, as shown in more detail in 2.3.3, the formal representation of collocational patterns which is made use of in ECD-type dictionaries (Mel'čuk et al. 1984; 1988; 1992; 1999), the system of lexical functions, constitutes a means to encode syntactic structure, together with semantic information. For instance, the function `Magn` is used to represent modifiers expressing intensification. Thus, when applied to a noun base (*smoker*), this function provides as its value an adjective (*heavy*), while, when applied to a verb (*to smoke*), the resulting value is an adverb (*heavily*).

Mel'čuk's approach is in accordance with what is proposed by authors working within the phraseological tradition. It has been mentioned above that the emphasis on classifying collocations according to their syntactic pattern was already emphasized in Palmer's (1933) work. Subsequently, Mitchell (1971, 52–53) proposed the concept of *colligation* to describe and classify combinations according to their syntactic pattern, e.g. adjective+agentive noun (*heavy drinker*). Nevertheless, it appears to be Hausmann's (1989, 1010) work which set the standards for the description of collocations in this respect, at least in the context of lexicography. Importantly, for Hausmann, the correspondence of a given combination to one of six specific syntactic patterns established by the author constitutes a defining criterion for it to be considered a collocation. This list of patterns, which includes noun+(epithet)adjective, noun(subject)+verb, verb+noun(object), verb+adverb, adjective+adverb and noun+(preposition)+noun combinations, was adopted and extended by subsequent authors. Another example of defining and classifying collocations with lexicographical purposes is constituted by the work of Benson et al. (1986b; 1986c), who, as we have seen, propose a detailed classification of lexical and grammatical collocations. The classification of collocations according to their syntactic pattern in the work of different authors as well as in collocation dictionaries is discussed in more detail in 2.3.1.

As explained in 2.2.3, according to Mel'čuk, the relationship between the elements of the collocation is directional, since the base lexically determines the selection of the collocate to express a given meaning. As it was already mentioned, Hausmann (1979, 191–192; 1989, 1010) conceives of directionality in the relationship of collocational elements in the same way. Following his description, the base is semantically autonomous, while the collocate is selected by the base, and receives its semantic identity only when part of the collocation. For instance, in *explode a myth*, the base *myth* chooses the collocate verb to express the meaning 'prove false'. A similar idea is conveyed by Mitchell (1971, 55), who describes restricted lexical selection in terms of "linguistic dependency". According to this author, in the expression *green as grass*, *green* is modified by *grass*, whereas in *green with envy, envy* is modified by *green*, such that, the modifying elements express the same meaning in both cases, that of intensification.

The way in which the above mentioned authors attribute a clear role to each of the two elements of the collocation contrasts with notions put forward by other scholars. We have seen that, in the frequency-based approach, both Halliday (1966, 156) and Sinclair (1966, 418) propose to analyze collocations in terms of probability of co-occurrence or significance, which implies determining the likeliness of one element co-occurring with the other through taking into account the number of individual occurrences and the instances of co-occurrence (see 2.2.1). Both authors noted that the notion of probability reflects an uneven relationship between the elements of a collocation, given the difference in their individual frequencies. More specifically, the probability for the less frequent element to occur in a given collocation is higher, as the collocation instances constitute a higher portion of the totality of its individual instances.

Note that while probability can be an indicator of restricted selection, it does not represent directionality in the same manner as it is described by Mel'čuk and Hausmann. For instance, the adjective *cerval* 'deer-like' co-occurs with a higher probability with *miedo* than the other way around, which implies that an instance of the adjective *cerval* predicts rather reliably an occurrence of the noun *miedo*. Nevertheless, for the latter authors, it is in fact the base *miedo* that selects the collocate *cerval* to express the meaning of intensification. On the contrary, in the case of the collocation *tener miedo* lit. 'to have fear' 'be afraid', the probability of the base co-occurring with the high frequency verb *tener* is higher, such that it both lexically selects the collocate verb, and is a better predictor of its occurrence than the other way around.

Commutability referring to the elements of the collocation (see e.g. Aisenstadt 1979; 1981; Cowie 1981; 1998a; Howarth 1996; 1998a; 1998b) is another non-reciprocal feature indicating restrictedness. However, similarly to the case of the likeliness of co-occurrence, it cannot be used to determine the directionality of restricted lexical selection in Mel'čuk's and Hausmann's sense. For instance, in the case of *make a decision*, used by Aisenstadt (1979, 73; 1981, 55–57) to exemplify restricted commutability, the noun *decision* can only be combined with a restricted number of verbs (*reach, take, arrive at, come to*) and constitutes the base of the collocation. On the contrary, the adjectives *auburn* and *hazel*, whose use is confined to given nouns, *hair* and *eyes*, respectively, are collocates, i.e. they are selected by these nouns, which in turn can co-occur with a variety of adjectives.

Finally, Bosque's (2001, 15–20; 2004a, XCV–XCVI) notion of collocation establishes directionality opposite to what we have seen in the case of the ECL, since, according to this author, it is the *collocate* or, following his terminology, the *predicate* that selects the *base* or *argument*. Restrictions on the base are of semantic nature, so that a collocate not only selects for a stand-alone base, but for groups of bases forming a *lexical class*, whose semantic features can be specified. For instance, the adverb *universalmente* 'universally', selects lexical classes of verbs expressing e.g. 'acceptance' (*aceptar* 'to accept', *admitir* 'to admit', *acoger* 'to embrace'), 'appreciation' (*aplaudir* 'to applaud', *celebrar* 'to celebrate', *reconocer* 'to recognize') and 'disapproval' (*condenar* 'to condemn', *detestar* 'to detest', *repudiar* 'to repudiate'). A similar idea is described by Cruse (1986, 104), who claims that in head-modifier constructions the *selector* element is the modifier given that it presupposes certain semantic traits of the head, the *selectee*, while in head-complement constructions the *selector* is the head and the complement is the *selectee*. Thus, according to this author, the adjective *pregnant* selects for a noun, or implies the presupposition that the modified noun has the trait "female", while the verb *drink* presupposes that its direct object has the trait "liquid".

We have seen that directionality in the relationship of collocational elements can be considered from different perspectives. Nevertheless, if we aim to describe collocations in a dictionary where the entry of one lexical item contains a list of collocating elements (as is the case of collocation dictionaries, see 3.4.2.1.B), from the point of view of language production, it is relatively straightforward to argue that the most suitable approach is the one proposed by Mel'čuk and Hausmann. When producing the combination *explode a myth*, the speaker is essentially aiming to convey the meaning 'prove false a myth', i.e. she is interested in finding the verbs the noun *myth* can be combined with to express the given meaning. In this context, it is less relevant whether the use of the verb *explode* is restricted to a small set of nouns which can be described along semantic traits, or whether an instance of the verb is a good predictor of an occurrence of the noun *myth.*

2.2.4.5 Commutability of components and restricted selection

Restricted commutability and restricted selection of the components are the traits that received most attention when it comes to describing collocations within the phraseological approach. They are, in fact, together with compositionality, discussed in the following section, the defining features used to distinguish between different types of phraseological expressions or phrasemes, as well as to distinguish between collocations and free combinations. Nevertheless, as it was already suggested when discussing directionality, authors differ in their ways of describing this notion.

As explained in 2.2.3, in his description of different types of phrasemes, Mel'čuk (2012, 32–42; 2013, 132–145) distinguishes between lexical, semantic and pragmatic restrictions. In the case of collocations, only lexical restrictions apply, namely on the choice of the collocate, which is determined by the base. Note that, contrary to approaches represented by other authors to be discussed below, Mel'čuk's notion of restrictedness is not particularly related to quantitative aspects of commutability, i.e. to whether a given element of a collocation can be substituted for many or few other elements – as it is found in the case of certain other authors (see below) –, but rather to unpredictability or arbitrariness. In addition, given that he conceives of lexical restrictions as the individual property of lexical items, Mel'čuk also differs from other authors in that he does not classify collocations according to degrees of restrictedness or define lexical restrictions in relation to semantic properties, although, – as it is discussed in 2.3.2.2 – he does admit to the fact that it is potentially possible to make generalizations over restrictions on a semantic basis (see Mel'čuk and Wanner 1996).

As it is pointed out by e.g. Alonso Ramos (2002, 67–68; 2006, 32), in order to understand the notion of collocation formulated within the ECL framework, these expressions should at all times be examined from the perspective of language production, and, especially, from a contrastive point of view. For a non-native speaker, both more idiosyncratic combinations such as *ignorancia supina* 'crass ignorance' and seemingly ordinary combinations such as *gran ignorancia* 'great ignorance' can result unpredictable. In this sense, collocations can be conceived of as combinations conveying meanings commonly associated to the base, which may not be obtained through the literal translation of a synonymous expression in another language. One such common meaning is that of intensification, which can typically be conveyed by more or less idiosyncratic adjectives associated to e.g. a noun expressing a cognitive state, as in the above examples involving *ignorancia* 'ignorance' (see Alonso Ramos 2006).

As mentioned in the previous section, restricted commutability, defined by Howarth (1996, 41) as the restrictions on the possible substitution of one component of a phraseological expression, is used by various authors working within the phraseological approach to describe the characteristics of phraseological expressions. The following paragraphs provide an overview of two aspects relating this notion. Firstly, I review how commutability or restricted selection is used to delimit different types of phraseological expressions, secondly, I provide an account of how collocations are further described or categorized through an analysis of the nature of commutability restrictions.

Most authors make use of the notion of commutability as a defining criterion which allows to delimit different types of combinations. For instance, Aisenstadt (1981, 54) claims that restricted collocations – which she contrasts with idioms and free combinations – are "combinations of two or more words, the components of which are used in one of their unidiomatic (often secondary, abstract, figurative) meanings, which follow certain structural patterns, and in which *one word at least is restricted in its commutability* not only by its grammatical and semantic valance, but also by usage" [emphasis mine]. In the same vain, for Hausmann (1989, 1010), it is restricted commutability or arbitrariness that distinguishes collocations, such as *feuilleter un livre* ‘leaf through a book’ from free combinations such as *acheter un livre* ‘buy a book’ (see also Benson 1989, 3–4).

As discussed in 2.2.2, the typology of word combinations described by Cowie (1981; 1988) and Howarth (1996; 1998a; 1998b) includes collocations together with idioms under the more general category called *composites* (see Table 1). Composites are word combinations with a “more or less invariable” form and a “more or less unitary” referential or propositional meaning, functioning below the sentence level, such as *spill the beans, dry run* and *do a U-turn* (Cowie 1988, 134–135). The criterion of commutability is said to distinguish collocations from idioms, since, while at least one component of the former generally allows substitution, the latter are immutable in what refers to the composing lexical elements (Cowie 1981, 227–228). Commutability is also used to differentiate between free and restricted collocations. Cowie (1981, 226) claims that free collocations (e.g. *run a business*, *explode a bomb*) show “openness of collocablity of each element in relation to other or others”, while restricted collocations (e.g. *command respect, escape someone’s attention*) are “characterized by extreme restriction on collocability”.

It should be noted, however, that the set of expressions termed *restricted collocation* by Cowie and Howarth, does not completely overlap with other authors’ use of the term, such as e.g. Aisenstadt’s. For the latter, it is solely restricted combinability that sets apart restricted and free combinations, while, for the former, restricted collocations, in addition to being restricted, obligatorily contain an element used in a figurative sense (see below). In addition, although both Cowie (1981, 226) and Howarth (1996, 34) liken their category of open/free collocations and those established by other authors, such as Aisenstadt’s free combinations and Mel’čuk’s free phrases (see 2.2.2 and 2.2.3), in my view, said categories are not completely equivalent. Free collocations, rather vaguely defined in Cowie’s and Howarth’s typology, seem to constitute a middle ground between what other authors term free combinations and collocations, since, in addition to nonce combinations, they also include expressions characterized by certain degree of arbitrariness (see Howarth 1996, 37), which in fact can be problematic for language learners (see Cowie 1981, 226).

Cowie (1981, 227) introduces the concept of *collocational range* to describe and quantify the set of lexical items which collocate with a given lexical element. He observes that, in the case of free collocations, collocational range can be determined through semantic properties of the set, i.e. *explode* in its literal sense co-occurs with 'explodable' nouns such as *bomb, mine,* etc. Note, however, that this is also true in the case of combinations, which, given the figurative meaning adopted by one of their elements, Cowie includes in the group of restricted collocations. For instance, *explode* in the figurative sense 'prove something false' co-occurs with nouns designating things which can be 'proven to be false' such as *myth, belief, theory.* Hence this interpretation of commutability appears somewhat problematic and cannot serve as a sole criterion to distinguish the so called free and restricted categories (see Howarth 1996, 42).

Bosque's (2001, 15–20) work represents a different understanding of the semantic properties of combinatory restrictions from that of Cowie's. In fact, he claims that Cowie's distinction between open and restricted collocations on the basis of restrictedness is irrelevant, since all collocations are restricted (Bosque 2011, ix), and goes on to describe collocations in function of the restrictions imposed by the collocate/predicate on the base/argument, claiming that bases that co-occur with a collocate can be perceived as forming paradigms called *lexical classes* (see 2.2.2 and 2.2.4.4). Hence, the defining criterion used to distinguish between collocations and free combinations is resolved in terms of whether the arguments co-occurring with a given predicate can be grouped into lexical classes on the basis of more or less concrete semantic criteria. For instance, Bosque sees no particular interest in including the adverb *lentamente* 'slowly' in a collocation dictionary, since it can co-occur with any verb denoting an action, while other adverbs, such as *energéticamente* 'energetically', *rotundamente* 'categorically' and *intensamente* 'intensely' co-occur with more restricted classes of verbs, thus form collocations.

In addition to being used as a definitorial feature, commutability restrictions are further analyzed in the work of some authors, and, in certain cases, give rise to typologies according to which collocations can be classified. Aisenstadt (1979, 73; 1981, 55–57) distinguishes between two types of collocations through taking into account the quantitative aspect of commutability, i.e. the number of different items a given lexical item can be combined with (see also Cowie 1981, 227–228). In the case of the first category, only one of the two elements is limited in commutability, such as in support verb+deverbal noun combinations. For instance, as we have seen above, the noun *decision* combines only with a limited set of verbs (*make*/*reach*/*take*/*arrive at*/*come to a decision*), while at least some of the verbs are less restricted in that they can co-occur with a large number of nouns. Another example is that of *auburn hair* and *hazel eyes*, where each adjective is restricted to the given noun, while the nouns can combine with other adjectives. In the case of the second category, both elements are restricted. For example, in *shrug one's shoulders* both the noun and the verb are restricted in commutability to a limited number of items: *square*/*hunch one's shoulders*.

Howarth (1996, 43) proposes a typology consisting of three degrees of restriction, taking into account not only commutability, but also the emergence of contextually restricted figurative meaning. Collocations in the first category are the most strictly restricted since none of the components can be substituted; in addition, the element with figurative meaning only acquires such meaning when occurring in the given combination, e.g. *curry favor.* The second group comprises combinations where the meaning of the figurative element can appear with a series of lexical items, e.g. *table a motion*/*bill*/*an amendment, adopt*/*assume*/*take on a role.* The third category includes the least restricted collocations, where both the element with figurative meaning and the literal element can be substituted, the latter with non-synonymous but semantically related items, e.g. *carry out*/*conduct*/*an experiment*/*a test*/*a survey.* This last type of combination is especially complex when it comes to describing collocational range, since often neither of the elements of the combination has a range which would entirely overlap with the other. Thus, e.g. one can *perform*/*carry out*/*conduct* both a *test* and an *experiment*, a *task* can only be *performed* or *carried out* but not **conducted*, while a *survey* is *carried out* or *conducted,* but not **performed* (Cowie 1998a, 3169; see also Howarth 1996, 44; 1998b, 37–39; 1998a, 173–175).

In his analysis of the semantic features of commutability restrictions, Cruse (1986, 278–280) establishes a distinction between logically necessary *selectional restrictions* and arbitrary *collocational restrictions.* In the first case, co-occurrence restrictions follow from the propositional meaning of a lexical item, while, in the second, one finds further restrictions not represented in the propositional meaning. For instance, the verb *to die* can co-occur with any animate subject, i.e. represents a case of logical restriction, while *to pass away* and the idiom *to kick the bucket* are only used with human subjects, e.g. ?*the daffodils kicked the bucket/died.* These last two cases, according to the author, represent examples of arbitrary selection, since the propositional meaning of both expressions corresponds to that of 'die' rather than 'die in a characteristically human way'. According to Cruse (1986, 279), collocational restrictions are not relevant to truth-conditions, meaning that a question such as “Have the daffodils kicked the bucket?” would not be answered with ‘No’ in the circumstances when the statement “The daffodils have died” is true, rather with ‘You can’t say that’.

Other authors provide a variety of further examples for the semantically non-motivated, arbitrary nature of commutability restrictions. Cowie (1994, 3169) highlights the existence of arbitrary combinatory restrictions in the case of *cut one’s throat, slash one’s wrist, *slash one’s throat* and ?*cut one’s wrist*, even though the elements of these collocations are used with their literal meanings. Mitchell (1971, 54) shows a case of arbitrary restriction affecting nouns designating occupations and verbs with the meaning ‘forced to terminate an occupation’, “operating in the cases of *barristers* who are *disbarred, doctors* who are *struck off,* […]”, etc. Bosque (2001, 22), to whose work commutability restrictions are central, provides, among others, the example of the adjective *diametralmente* ‘diametrically’ which can be combined with the noun *opuesto* ‘opposed’ but not with *distinto* ‘distinct’ or *contrario* ‘opposite’. Lastly, another, often cited example of arbitrary restrictions is the use of support verbs, having little, if any, meaning when combined with a corresponding noun, however highly constrained in their use, e.g. *take a look, *make a look, assume importance, *adopt importance* (Cowie 1994, 3169; Howarth 1996, 40–41).

Cruse (1986, 280–282) describes three subtypes of collocational restrictions according to the degree to which required semantic traits can be specified for in the case of each. *Systematic collocational restrictions* are fully specifiable. The example provided by the author is that of the verbs *grill* and *toast,* which denote the same action, with the difference that food items grilled are raw, while those toasted are already cooked. The term *semi-systematic collocational restrictions* is used for cases when required semantic traits show a certain tendency, but with exceptions to it, such as with the noun *costumer,* typically used to refer to a person who acquires something for money, contrary to *client,* typically a person receiving professional or technical service. An exception to this, mentioned by Cruse, is the case a person using the services of a bank, who can be called a costumer. The third type, *idiosyncratic collocational restriction*, refers to cases when the collocational range of a lexical item can only be specified thorough listing the items it can co-occur with, for instance, the adjective *flawless* can collocate with the nouns *performance, argument* and *complexion,* but not with *behavior* or *reputation.*

Finally, similarly to Cruse, Bosque (2011, ix–x), also examines commutability restrictions from a semantic perspective, and establishes the categories of *deducible* and *non-deducible collocations*. In the case of deducible collocations, one can identify extensive paradigms of bases which combine with a given collocate on semantic grounds. For example, the adverb *substantially* combines with e.g. 'verbs of increasing', such as *enhance, improve, progress,* etc., 'verbs of decreasing', like *decline, fade, reduce, tighten,* etc., and 'verbs denoting change', as e.g. *alter, amend, modify* or *rewrite*. On the contrary, in the case of non-deducible collocations, bases constitute closed paradigms, which can be limited to a single element. For instance, the verb *bleat* can only have *sheep* or *goat* as possible subjects, while the verb in the Spanish collocation *conciliar el sueño* 'to get to sleep' appears with the given meaning only in this particular combination. According to the author, these two categories form part of the native speaker's competence.

To round up this rather lengthy discussion of restricted selection or commutability, I would like to emphasize the difference between the notion as outlined in Mel'čuk's work and that of other authors. Firstly, in the case of the ECL framework, lexical selection is conceived of as strictly directional, with the base of the collocation determining the choice of the collocate. Secondly, lexical restrictedness is related to the unpredictability or arbitrariness of lexical selection determined by the base, hence it is perhaps more accurately described by the term restricted selection, than conceived of in more strictly quantitative terms of commutability, which – as we have seen - is not necessarily understood as a directional property. Thirdly, lexical restrictions in Mel'čuk's terms are the individual property of the lexical item constituting the base, and are to be described in its lexical entry, contrarily to the view expressed by e.g. Bosque, according to whom combinatorial restrictions can be systematized through describing the semantic properties of bases combining with a given collocate.

2.2.4.6 Compositionality and semantic transparence

Besides restricted commutability, compositionality is the other main trait used to delimit the category of collocations, differentiating these expressions from idioms. Authors belonging to the phraseological tradition agree that while collocations (e.g. *empty a bucket*) have a compositional meaning, the meaning of idioms (e.g. *kick the bucket*) can be characterized as unitary or non-compositional. In other words, in the case of collocations, the meaning of the expression is provided by the meanings of its elements, contrary to the case of idioms, whose meaning is linked to the expression as a whole (see e.g. Aisenstadt 1979; 1981; Cowie 1981; Cruse 1986, 40–41; Hausmann 1989; Howarth 1996; Mel'čuk 2012; 2013, 132–133).

While collocations are considered to have a compositional meaning, most authors further explore the ways their components contribute to the meaning of the whole expression, and describe different subcategories or types of collocations. We have seen that Mel'čuk, as well as Hausmann (1998, 65–66) distinguish between the two components of a collocation, the base and the collocate. As for its semantic properties, the base is described as freely selected by the speaker according to its meaning, in other words it is an *auto-semantic* or semantically 'self-sufficient' element. The collocate is considered to be chosen depending on the base, thus its meaning should always be defined relative to the context; it is, in Hausmann's words, a *non-semantic* ("sinsemántico") element. An extreme case is that of Spanish *solo* lit. 'alone', which when combined with the noun *café* 'coffee' obtains the meaning 'without milk', not found in any other combination. As mentioned in 2.2, this notion of contextual determination of meaning in the case of the collocate was introduced by the Russian phraseologists. A description of the same process is provided by Allerton (1984, 23; 1982, 27–29) under the term *semantic tailoring*, while the semantic characteristics of the collocate are further explored by Alonso Ramos (2002, 83–86; 2012, 20–22), who discusses whether they have a sufficiently autonomous meaning so as to be described in individual dictionary entries.

Mel'čuk (1998, 29–30) distinguishes four major types of collocation according to the semantic contribution of the collocate to the meaning of the combination: 1) In the first type, the collocate does not add any semantic content to the expression, rather functions as a semi-auxiliary which has been selected by the base to support it in the given syntactic configuration. Such is the case of support (or light) verb constructions, e.g. *do a favor, take a step, launch an appeal*. 2) In the second type, the collocate acquires a meaning which it only has in combination with the base, and maybe with a few other lexemes. An example is *black coffee*, where black acquires the meaning 'without milk'. 3) In the case of the third type, the collocate expresses the same meaning it would normally have in other contexts, however its synonyms cannot always be combined with the given base, e.g. *strong/*powerful coffee, heavy/*weighty smoker*. 4) Finally, in the fourth type of collocation, the collocate again expresses its regular meaning, and, in addition, its meaning is closely related to that of the base, in other words, the collocate can be said to be “bound” by the base. Such collocations are e.g. *the horse neighs, rancid butter,* and *artesian well*.

In the case of Cowie (1981, 226) and Howarth (1996, 38), as we have seen, the description of the semantic features of collocational elements is related to the distinction between free and restricted collocations. In free collocations, such as *to cut bread* or *to eat cheese,* each element is used in its primary, literal sense, while in the case of restricted collocations, e.g. *to foot the bill,* one element (*foot*) has a figurative sense. Although the authors do not define explicitly what they mean by figurative meaning, since the difference between free and restricted collocations is also described in terms of collocational range, it can be assumed that the term figurative refers to a meaning that is only found in a limited context. This seems to be somewhat problematic, as, for instance, Cowie (1981, 226) argues that *run* in the sense of ‘manage’, in e.g. *run a business,* represents a literal meaning, while he also claims that the verb *manage* has a more restricted collocational range than *run*, nevertheless, nobody would argue that the use of this verb in e.g. *manage a bank* is not literal. At the same time, as it was already discussed in 2.2.4.5, the use of the notion of collocational range for differentiating between free and restricted collocations is also questionable, since combinations can be semantically motivated regardless of the commutability of their elements in quantitative terms, as we have seen in the case of literal and figurative *explode,* as well as transitive *run* (see also Bosque 2011, xiv).

Aisenstadt (1979, 73; 1981, 57–59; see also Corpas Pastor 1996, 83) establishes three categories in order to describe the degree of semantic specialization of the element of the collocation with contextually determined figurative meaning. These categories, as we have seen above in the case of Mel’čuk, establish a relationship between semantic specialization and commutability restrictions. In the case of collocations belonging to the first group, the figurative element has a narrow, specific meaning, limiting the commutability of the given lexical item, such as *shrug* ‘move one’s shoulders up and let them drop’ in *shrug one’s shoulder, foot* ‘pay’ in *foot the bill* or *fruncir* ‘move one’s eyebrows closer together’ in *fruncir el ceño* ‘to knit one’s brow’. In the case of the second group, one lexical element is used in a secondary, abstract or figurative meaning with restricted commutability, which contrasts with a concrete meaning of the same item used in other – less restricted – contexts. Examples for the use of lexical items with such secondary meaning are the verb *pay* in *pay attention*/*heed*/*respects,* etc. or *sofocar* ‘to suffocate' in *sofocar una revuelta* 'to put down a riot'. The third category established by Aisenstadt involves combinations containing a grammaticalized element with a delexicalized, vague meaning, such as support verb combinations, e.g. *give a laugh, have a fall, dar comienzo* lit. 'to give start'.

Commutability, as well as the figurative or specialized nature of the meaning of the collocate is often considered to be related to semantic transparence or opacity. For instance, Koike (2002, 7) maintains that the semantic transparence of collocations is variable and that it depends on the degree of semantic specialization of their elements. Combinations with both elements used in their direct sense are fully compositional and fully transparent (*abrir la ventana* 'to open the window'). In contrast, in cases when one element is used in a figurative sense (*ahuyentar el temor* 'to chase away the fear'), the meaning can be transparent or deducible to some extent given its similitude to literal uses (*ahuyentar un perro* 'chase away a dog'), although the more metaphorical this figurative use becomes, the more opaque the meaning of the combination (*amasar una fortuna* 'to amass a fortune', lit. to knead a fortune'). Collocations with both elements used in a figurative or metaphorical sense, have an even more opaque meaning, and are rather similar to idioms (e.g. *cortar la palabra* lit. 'cut sb's word' 'interrupt sb's speech'). Importantly, according to Koike, as well as Howarth (1996, 44), the semantic opacity of certain combinations affects both encoding and decoding.

On the contrary, as it was already mentioned, Mel'čuk (2013, 133), emphasizes that opacity or transparence are subjective and gradable properties, claiming, together with Alonso Ramos (2006, 32–33), that it is not a defining property of collocations, and should only be considered from the perspective of reception. While collocations can be more or less transparent depending on e.g. the mother tongue of a language learner or whether he or she is familiar with the meanings of the components of a given combination, from the point of view of production, all collocations should be considered to be lexically restricted. In a similar vein, Hausmann (1979, 191–192; 1989, 1010; 1998) also emphasizes that, while restricted lexical selection potentially hinders the production of collocations in the case of a language learner (see 2.2.4.7), it does not interfere with their transparence in decoding, i.e. restrictedness and transparence should not be correlated. The way in which the different characteristics of collocations contribute to their problematic nature when it comes to learning a foreign language is discussed in more detail in the following section.

2.2.4.7 Learning difficulties characterizing collocations

The last aspect to be discussed involves the difficulties posed by collocations to the foreign language learner. As it was mentioned above, the study and description of multiword units, among them collocations, was to a great extant motivated by pedagogical purposes within the phraseological tradition. Consequently, this approach can be conceived of as a descriptive framework which provides a theoretical background to the creation of reference works dealing with different types of multiword units (see Gyllstad 2007, 6). Accordingly, authors ascribed to it often describe the main characteristics of phraseological expressions in relation to the problems they pose to the language learner.

The difficulties implied by collocations in the language learning process are rooted in the features of restrictedness and opacity. Authors' opinions diverge, as we have seen, both when it comes to defining the directionality of lexical restrictions, and concerning the degree to which restrictiveness and opacity affect language production, comprehension or both. Importantly, the perceived nature of difficulties is not only relevant for the way collocation is defined, but it also has a direct impact on the proposed design of lexicographical resources and teaching materials.

Although Mel'čuk (e.g. 1998; 2008; 2012) does not essentially approach the issue of restricted lexical selection from a pedagogical point of view, he does define collocation, similarly to Hausmann (1979, 191–192; 1989, 1010; 1998), as an eminently production phenomenon. The particularity of collocations is described as the speaker freely selecting a given lexical item, the base, according to its meaning, independently of other lexical items, while the selection of a second item, the collocate is controlled by the base. For instance, a speaker of Spanish can freely opt for using the noun *miedo* 'fear' in order to express the meaning 'X perceives Y as dangerous or threatening', regardless of what other lexical items have been selected. This is the case of *semantically controlled* lexical selection. In contrast, when the speaker intends to express 'start feeling' in relation to *miedo,* they have to take into account the noun when choosing the appropriate verb, e.g. *coger* 'to catch'. This second type of lexical choice is said to be both semantically and *lexically controlled.*

In accordance with this, Hausmann claims that a learner can acquire the bases individually, since they are used as autonomous elements, i.e. with their literal meaning. On the contrary, collocates should be inevitably learned as part of the given combination. This author also emphasizes that the structure of collocation dictionaries should respond to this learner need. Consequently, it is more convenient to describe collocates in the entry of the base, supporting language production, than to include a list of bases is the entry of a collocate, which would only serve to verify whether a collocation is correct or not, without allowing to look up unknown collocations.

Bosque (2011, xi) represents an opposing view regarding the structure of collocation dictionaries, as he proposes to present combinatorial information through providing lexical entries for collocates, where lists of corresponding lexical classes of bases are specified, thus aiding the acquisition process. This idea, implemented in the combinatorial dictionary *Redes* (Bosque 2004b), is derived from the author's interpretation of the directionality of lexical selection in the case of the components of collocations. As it was already explained in 2.2.4.4, according to Bosque it is the collocate that imposes restrictions on the selection of bases, or, rather groups of bases having common semantic traits, he refers to as lexical classes. As a consequence, he conceives of collocational knowledge as structured lists representing paradigms of bases that can be combined with a given collocate, and also assumes that collocations are in general not to be memorized individually, but in groups corresponding to lexical classes. The layout of lexical entries in collocation dictionaries will be further discussed and exemplified in 3.4.2.1.B.

As it was suggested in 2.2.4.6, unlike Mel'čuk and Hausmann, other authors attribute relative importance to reception problems, or else, problems of decoding resulting from the opacity of collocations. For instance, Howarth (1996, 44) maintains that opaque combinations, such as *foot the bill*, may constitute a principally decoding problem to the learner, and be less problematic from the point of view of production. The reason for this is that this type of combination is more marked and idiosyncratic, consequently, it is more salient to the learner, who, once having it acquired, will be unlikely to recombine an element with contextually bound meaning with a different lexical item, e.g. **foot the entrance fee.* On the contrary, semantically transparent collocations, e.g. *assume importance, adopt a role,* found toward the free end of the collocational spectrum, are more problematic in language production. This argument is supported by the claim that learners might be unaware of the arbitrary restrictions applying in the case of these apparently regular combinations.

Similarly to Howarth, García Platero (2002, 28) also perceives collocations as problematic both in encoding and decoding. He highlights the fact that most authors tend to neglect the second problem, claiming that familiarity with the individual meaning of each element composing the combination guarantees the correct interpretation. However, the author notes that, in certain cases, figurative meanings of collocates may not be sufficiently described in lexicographical works, hindering decoding. Such is the case of e.g. the combination *conciliar el sueño* ‘to get to sleep, lit. to reconcile sleep’, when the particular meaning of the verb is missing from the dictionary. Note that, this case, although clearly representing a difficulty for learners, is not necessarily brought forward by the nature of collocation, rather, it can be seen as a problem to be tackled in the lexicographic description of the collocate. In this respect Alonso Ramos’ (2011) emphasizes that figurative meanings of collocates should be described in a general monolingual or bilingual dictionary as separate meanings corresponding to a separate lexical unit in the entry of a given lemma, in the same way as literal meanings.

Before concluding this subsection, it should be noted that authors working more strictly in the field of foreign language teaching tend to propose a simplification or adaptation of the definition of collocation, so that, instead of emphasizing accuracy of linguistic description, they prefer to focus on learner needs (Gyllstad 2007; Ferrando Aramo 2009). Consequently, in this approach, the concept of collocation is often extended as compared to the use of the term in phraseology or lexicography, so that it is not limited to restricted combinations, but also includes non-restricted frequent combinations. Thus, the focus is rather on the usefulness of target expressions than on whether they fulfill certain linguistic criteria (Higueras García 2006:18-19). Despite of this, as it has been made clear along this chapter, the notion of collocation adopted in the present thesis is aimed to be as theoretically well founded and well-delimited as possible.

2.2.5 The concept of collocation in the present thesis

The present chapter has so far been concerned with discussing the phenomenon of collocation, as it is conceived of in the work of different authors, belonging to what were denoted as the frequency-based and the phraseological approach. I presented in detail the criteria used to establish the taxonomy of phrasemes within the ECL framework, whose notion of collocation is adopted in this thesis, and I compared the main characteristics of collocations as described within the ECL approach to what is outlined in the work of other authors. In order to conclude this discussion, I believe it is convenient to take stock of the main traits summarizing the concept of collocation used in this thesis.

The main feature characterizing collocations is restricted lexical selection, which involves one element of the collocation, the base, determining the selection of the combining element, the collocate, to express a given meaning. For instance, the Spanish noun *hábito* ‘habit’ chooses the collocate verbs *abandoner* ‘abandon’ or *dejar* ‘leave’ in order to express the meaning ‘stop having’. In ECL, restrictedness is not understood in terms of quantitative aspects of commutability, that is to say, it is not taken into account whether each of the components of the collocation can be substituted by few or many other lexical items. Rather, the emphasis is on the unpredictability or arbitrariness in the choice of the collocate, observed from the perspective of language production. Arbitrariness is most often judged from a contrastive perspective, which means that even apparently trivial or transparent combinations can be considered collocations, if their literal translation does not necessarily constitute a native-like combination in other languages. In order to identify these potentially arbitrary combinations, collocations can be conceived of as conveying meanings commonly associated to the base. The above example illustrates one such meaning related to *hábito* ‘habit’, which in English could in fact be expressed using the verb *abandon,* as well as *break, kick* or *give up* but not **leave.*

As for formal properties, as follows from the above, collocations are constituted by two elements, and are described as combinations of lexical units – not word forms or roots. The base and the collocate are syntactically related, such that it is possible to identify in the case of each language a set of prototypical syntactic patterns corresponding to collocations. When it comes to semantic properties, collocations have a compositional meaning, in the sense that, at all times, it is possible to identify a meaning component corresponding to the base and a meaning component corresponding to the collocate, although the meaning of the collocate may be specialized and bound to a limited number of combinations, such as in the case of *café solo* ‘black coffee’.

Finally, a few remarks are in place regarding frequency of co-occurrence and learning difficulties. We have seen that, in the ECL framework, frequency of co-occurrence has no role whatsoever in defining collocations. Accordingly, in the present thesis frequency is not used to determine whether a combination constitutes a collocation or not. We will see, however, that frequency information is made use of in collocation learning resources, such as collocation dictionaries and CALL resources, which will be described in the following chapters. Corpus frequency is also used to determine the correctness of learner collocations in study described in Chapter 4.

Regarding learning difficulties, it has been discussed that the ECL approach considers collocations strictly as production phenomena, given they are defined on the basis of the lexical restriction in the selection of the collocate. When it comes to lexical resources, this implies that collocation dictionaries should be oriented to production through listing possible collocates in the lexical entries of bases. Although I concur with this view, I also believe that in the age of electronic dictionaries, which potentially allow flexible searches throughout a lexical database, for the sake of practicality, it should also be possible to obtain information corresponding to what is described as decoding lexical entries by e.g. Alonso Ramos (2002, 86–93) This idea is reflected in the collocation learning tool presented in Chapter 6.

2.3 Collocation typologies

Collocations are classified according to various criteria depending on the specific aspect of the combinations studied. Previously, we saw that authors have grouped collocations according to the degree of restrictedness or commutability displayed by their components, semantic transparence, or else, the specificity of meaning of the collocate. Two aspects of the description and classification of collocations are especially relevant from the point of view of foreign language teaching and learning. These aspects refer to the syntactic pattern and the meaning of combinations, parameters used for organizing and presenting collocations in lexicographical works – and sometimes in teaching materials.

In what follows, Sections 2.3.1 and 2.3.2 concentrate on these two parameters as applied explicitly or implicitly for classifying collocations in descriptive studies as well as collocation dictionaries aimed principally at language learners. Section 2.3.3 introduces the description and classification of collocations proposed within the Meaning⇔Text Theory and used in different dictionaries following the ECL model, which, as we will see, consists of representing the semantic and syntactic properties of combinations through the system of lexical functions.

2.3.1 Classifying collocations according to their syntactic pattern

As it was explained in 2.2.4.4, some researchers whose work can be ascribed to the frequency-based approach to collocations (see Greenbaum 1970; Greenbaum 1974; Kilgarriff and Tugwell 2001; Kjellmer 1984; Mitchell 1971; Williams 1998) make use of pre-determined sets of syntactic patterns in order to single out combinations which have more interest in linguistic analysis. We have also seen that there is considerable emphasis on providing a detailed description of the syntactic pattern of collocations within the phraseological approach, starting already form Palmer's (1933) work (see Cowie 1998a), to the point that Hausmann (1989, 1010) explicitly defines collocations with reference to a definite set of patterns. In what follows, I present an overview of syntactic classifications of collocations provided by different authors, as well as typologies adopted by different collocation dictionaries, which in fact tend to use syntactic pattern as the main organizing principle within lexical entries.

To begin with, it should be noted, that a distinction is often made between so called *grammatical* and *lexical collocations* in the literature. This can be traced back to the *BBI dictionary of English word combinations* (BBI, Benson et al. 1986b), where grammatical collocations are described as consisting of a dominant element being a noun, adjective, verb, etc. and a preposition or a grammatical construction (p. xvi), while lexical collocations are characterized as consisting of nouns, adjectives, verbs and adverbs, and normally not containing prepositions, infinitives or clauses (p. xxx). The theoretical framework adopted by the present thesis does not acknowledge such distinction, since all collocations are conceived of as lexically restricted combinations, while most of what Benson et al. (1986b) define as grammatical collocations are not considered to be collocations at all, but are described as representing other types of linguistic phenomena.

According to the argument offered by Alonso Ramos (1993, 157) and Wanner (1996, 18–20), the great majority of grammatical collocations are to be described in the government pattern of a specific lexical unit. This is so, since e.g. a preposition governed by a verb – whose selection is indeed restricted, that is, the preposition is obligatorily required by the verb – does not add any semantic content to the expression, it is merely a grammatical element that serves to join the verb with its argument. Following this idea, from among the list of Benson et al.'s (1986b) eight types of English grammatical collocations[8], only one qualifies as lexically restricted combination. This is the case of

[8] The BBI distinguishes a total number of eight types of English grammatical collocations. The first seven of these contain a noun or an adjective as the dominant element: 1) noun+preposition (*blockade against*), 2) noun+*to*+infinitive (*be a pleasure to do sg*), 3) noun+*that*-clause (*reach an agreement that*), 4) preposition+noun (*by accident*), adjective+preposition (*be angry at sb*), 6) predicate adjective+*to*+infinitive (*be necessary to do sg*), 7) adjective+that clause (*be afraid that*). The eighth type of grammatical collocation includes nineteen subtypes of verbal patterns, such as e.g. verbs which can or cannot undergo dative movement transformation (*send sth to sb – send sb sth*; *describe sth to sb – *describe sb sth*),

preposition+noun combinations, such as e.g. *in anger* as in *raise one's voice in anger*, given that here the preposition does have semantic content, which might be paraphrased as 'showing or feeling' or 'as a demonstration of' (Wanner 1996, 20).

Note that the fact that combinations representing government pattern are not described as collocations in the framework of the present thesis, does not mean that they do not constitute useful information for a language learner. Some commercial collocation dictionaries, e.g. the *Oxford Collocations Dictionary* (OCD, McIntosh, Francis, and Poole 2009), the *Macmillan Collocations Dictionary* (MCD, Rundell 2010), and the *Longman Collocations Dictionary and Thesaurus* (LCD, Mayor 2013), in fact, do contain such combinations. For instance, in noun entries of the OCD we find combinations of the type noun+preposition (*hazard for*), while verb entries contain verb+preposition (*disagree with*) and adjective entries show adjective+preposition (*friendly towards*) combinations.

Following Hausmann's (1979; 1989) notion of collocation, adopted in the ECL framework, the following overview will organize possible syntactic patterns of collocations according to the part of speech of the base[9] (for a summary see Table 3). The syntactic patterns enumerated correspond to those normally described within the ECL framework, and, consequently, to the patterns taken into account in the empirical studies described in this thesis. Nevertheless, I also discuss additional types of combinations which are described in typologies underlying the design of the four English collocation dictionaries mentioned above, that is the BBI, the OCD, the MCD and the LCD; the work of Corpas Pastor (1996, 66–76), whose description of syntactic patterns of Spanish collocations follows Benson et al.'s (1986b; 1986c) and Hausmann's (1989) proposal, and was in turn adopted by Castillo Carballo (1998, 53–54); as well as Koike's (2001, 44–60) classification of Spanish collocations; and the typologies underlying the Spanish combinatory dictionaries *Redes* (Bosque 2004b), *Diccionario combinatorio práctico del*

verb+preposition (*consist of, adhere to*), verb+*to*+infinitive (*continue to do sg*), verb+possessive+gerund (*excuse sb's doing sg*) etc.

[9] Note that neither do all of the authors mentioned here adopt the approach of distinguishing between the base or the collocate, nor, as it will be discussed in more detail in 3.4.2.1.B, all combinatory dictionaries are necessarily consistent in presenting bases as headwords. These factors, however, are not taken into consideration in this account on collocational syntactic patterns.

español contemporáneo (*Práctico*, Bosque 2006) and *Diccionario de colocaciones del español* (DiCE, Alonso Ramos 2004) [10].

2.3.1.1 Collocations with a noun base

As for collocations with a noun functioning as the base, five main patterns may be identified, although only the first four of them are taken into account in the empirical studies forming part of this thesis: (1) NOUN$_{SUBJ}$+VERB, (2) VERB+NOUN$_{COMP,}$ (3) MODIFIER+NOUN, (4) NOUN+*of*/*de*+NOUN, (5) PREPOSITION+NOUN.

(1)(2) The typologies of all authors and all of the dictionaries mentioned above contain combinations corresponding to the NOUN$_{SUBJ}$+VERB (*la ira desvancece* 'sb's anger falls') and VERB+NOUN$_{COMP}$ patterns (*sacar dinero* 'withdraw money'), with the difference that the LCD and the Spanish combinatory dictionary, *Práctico* include all verbal collocates in the same group, making no explicit distinction. The BBI, the MCD, Corpas Pastor and Koike introduce a third verbal-nominal pattern, whereby the noun constitutes part of a prepositional phrase complementing the verb, as in *poner a prueba* 'put to trial'. At the same time, Penadés Martínez (2001, 68–69) observes that, since both in the case of Spanish verb+noun and verb+prep+noun collocations the noun is generally preceded by an article or a possessive pronoun, these cases can be more precisely represented through using the schemes V+NP and V+PP, respectively. Nevertheless, this thesis, in accordance with the categories customarily used in the ECL framework, only distinguishes verbal-nominal combinations regarding whether the base constitutes the subject or one of the complements of the noun.

(3) The MODIFIER+NOUN pattern (note that, in most cases, the default word order in Spanish is the opposite: NOUN+MODIFIER) corresponds to two main subtypes generally mentioned in the literature. The first of these, collocations containing a noun modified by an adjective (*vino tinto* 'red wine') is included in all typologies. The status of the second, where the base noun co-occurs with a noun functioning as an attributive modifier (*decisión clave* 'key decision' or *cama nido* 'trundle bed') is more debated. It is included in the four English collocation dictionaries, as well as in *Práctico* and Corpas Pastors' typology, while, Koike (2001, 44–60) claims that noun+noun combinations (further examples are *paquete bomba* 'parcel bomb' *hombre clave* 'key person', *ciudad fantasma* 'ghost town')

[10] This dictionary contains collocations of Spanish emotion nouns, hence collocation patterns included are limited to collocations which have a noun functioning as base.

constitute compounds. For more details regarding the debate as to the status of noun+noun combinations – in the specific case of Spanish – one may refer to Val Álvaro (1999) arguing for these combinations to be considered nominal compounds or Suñer Gratacós (1999) claiming that they represent cases of apposition. For the purposes of the present thesis, this type of combination will be considered a collocation in all cases when the criterion of compositionality is fulfilled.

(4) The pattern NOUN+*of*+NOUN (*bouquet of flowers, bar of chocolate*), the Spanish equivalent of which is NOUN+*de*+NOUN (*rebaño de obejas* 'flock of sheep' *onza de chocolate* 'square of chocolate') is included in the BBI and the OCD, as well as in Corpas Pastor's and Koike's typology, *Práctico* and DiCE. While the BBI specifies that the first noun implied in the collocation is necessarily a quantifier (see the examples above), the dictionaries *Práctico* and DiCE also include combinations in which the semantic relationship between the two nouns is not necessarily that of quantification (*prueba de amor* 'demonstration of one's love', *acceso de remordimiento* 'fit of remorse', *señal de remordimiento* 'sign of remorse'). The OCD distinguishes between quantifier+*of*+noun combinations and other types of NOUN+*of*+NOUN collocations through including the latter under the heading "phrases".

(5) As it was discussed above, combinations of the type PREPOSITION+NOUN (*entre dificultades* 'in hardship', *de casualidad* 'by accident'), included in the category of grammatical collocations in the BBI, are considered to be lexically restricted combinations in the ECL framework. Combinations of this type are included in *Práctico*, as well as in the OCD and LCD, in the latter two cases, under the heading 'phrases'. In DiCE, this type of combination is included in the group of *participant attributes*, containing both combinations and single word expressions making reference to the participants of the situation designated by the lexical unit constituting the base. For instance, *con enfado* 'with anger' is used to describe the behavior of a person experiencing anger. Note, however, that this last collocation pattern is not taken into account in the empirical studies described in this thesis.

In the case of the noun entries of the MCD, we find a number of further combination-types; these, however, are not treated as collocations within the ECL framework. One such type of combination is constituted by two coordinated nouns with invariable word order (*alcohol and gambling, goods and services*); another combination-type corresponds to two nouns joined by a preposition other than *of* (*immunity against* *infection*), which can be interpreted as frequent combinations generally exemplifying government pattern.

2.3.1.2 Collocations with a verb base

(1) The only collocation pattern with a verb base is VERB+ADVERB (*rain heavily, apologize humbly, negar rotundamente* 'deny categorically'). This pattern can be found in all of the above mentioned typologies and dictionaries, while the semantic and combinatorial features of Spanish VERB+ADVERB collocations are studied in detail by García-Page (2001). Note that, as a result of her study of a list of Spanish collocations, Penadés Martínez (2001, 69) proposes a further collocation pattern, where the verb is followed by a comparative construction, as in *dormir como un tronco* 'to sleep as a log'. Although this constitutes a recurrent pattern in Spanish, from the point of view of the theoretical framework followed here, *como un tronco* is more precisely described as an idiom filling the position of an adverb, thus combinations analogous to the one exemplified here are considered to correspond to the pattern VERB+ADVERB.

The collocation dictionaries OCD, MCD, LCD and *Práctico* also distinguish a type of collocation constituted by the combination of two verbs (*fail to detect, need to detain, seek to illustrate*). According to the user instructions provided in *Práctico* (Bosque 2006, XLIX), the dictionary includes three types of such combinations: the first group is constituted by semi-periphrastic expressions, like *echarse a llorar* 'start crying'; the second type contains combinations which are not precisely semi-idiomatic, but where the argument is subject to semantic selection by the predicate, such as *sacar a bailar* lit. 'lead to the dance floor, lit. take out to dance'; and the third type is described as containing useful, relevant, or frequent expressions, like *eludir pronunciarse* 'avoid expressing one's opinion'. In the BBI, the same types of combinations are classified as grammatical collocations, through indicating in the lexical entry of e.g. the verbs *fail, need* and *seek* that they are followed by a *to* + infinitive complement. As mentioned earlier, in the framework of this study, this phenomenon is described in the government pattern of the corresponding lexical items, and is not considered to constitute restricted lexical choice.

Furthermore, as in the case of nouns, the MCD includes in its verb entries combinations of coordinated verbs (*relax and unwind, inspire and motivate*), which are frequent and tend to have a fixed word order.

Finally, the Spanish combinatory dictionary *Práctico* describes a further type of collocation where the verb is preceded by a governing preposition. An example provided for the case of this pattern is *sin vacilar* 'without hesitating'. Note, however, that, while this particular combination does not show lexical restriction, the pattern itself is not consistently represented throughout the dictionary, since homologous expressions such as *sin pensar* 'without thinking, considering' are missing from its nomenclature.

2.3.1.3 Collocations with an adjective base

The two collocation patterns with an adjective base identified are the following: (1) MODIFIER+ADJECTIVE and (2) VERB+ADJECTIVE.

(1) Collocations consisting of an adjective and a modifying adverb (*seriously injured, widely excessive, estrechamente ligado* 'closely bound') are included in the typologies provided by all the authors and dictionaries mentioned at the beginning of this discussion. Note that collocations involving colors, such as *pale green,* are described as adjective+adjective combinations in the MCD and the LCD, while in e.g. *Práctico,* names of colors are treated as nouns, such that *azul marino* 'navy blue' is described as a noun+adjective combination.

(2) The less commonly treated collocation pattern, VERB+ADJECTIVE (*prove disappointing, become desirable, salir malparado* 'come off badly') included in the OCD, MCD and in Koike's typology. Here we find a verb, functioning as a copula, whose selection is determined by the adjective in use.

The MCD includes a series of other types of combination in its adjective entries, which, as in previous cases, seem to constitute frequent or useful expressions, or represent the government pattern of the headword, without being genuine cases of lexical restriction. Combinations of an adjective with a verb in the infinitive form, such as the case of *glad to hear*, despite being frequent co-occurrences, exemplify the government pattern of the adjective. Similarly, in the case of adjective+preposition+noun combinations (*generous with time*), the more relevant linguistic information is constituted by the governed preposition selected by the adjective. Finally, combinations of coordinated adjectives (*desolate and lonely*), in the same way as combinations consisting of coordinated nouns and verbs mentioned before, rather than collocations, can be considered frequent or routinely expressions.

2.3.1.4 Summary of collocational syntactic patterns

The aforementioned syntactic patterns of collocations are summarized in Table 3, together with the collocation typologies or collocation dictionaries in which they are mentioned.

Base	**Pattern**		**Example**	**Hausmann**	**BBI**	**OCD**	**MCD**	**LCD**	**Corpas**	**Koike**	***Práctico***	**DiCE**
N	**N$_{\text{Subj}}$+V**		*la ira desvancece* 'sb's anger falls'	+	+	+	+	+	+	+	+	+
	V+N$_{\text{Comp}}$	V+N	*sacar dinero* 'withraw money'	+	+	+	+	+	+	+	+	+
		V+prep+N	*poner a prueba* 'put to trial'	-	+	+	+	+	+	+	+	+
	Modif+N	Adj+N	*vino tinto* 'red wine'	+	+	+	+	+	+	+	+	+
		N+N	*ciudad fantasma* 'ghost town'	-	+	+	+	+	+	-	+	-
	N+prep+N		*ramo de flores* 'bouquet of flowers'	+	+	+		(+)	+	+	+	+
	Prep+N		*entre dificultades* 'in hardship'	-	Gram.	+	-	+	-	-	+	+
V	**V+Adv**		*negar rotundamente* 'deny categorically'	+	+	+	+	+	+	+	+	-
Adj	**Modif+Adj**	Adv+Adj	*estrechamente ligado* 'closely bound'	+	+	+	+	+	+	+	+	-
	V+Adj		*salir malparado* 'come off badly'	-	-	+	+	+	-	-	-	-

Table 3 Summary of collocational syntactic patterns

2.3.2 Classifying collocations according to their meaning

As it is attested by the previous sections, most descriptive studies and all lexicographic works classify collocations according to their syntactic patterns. In addition, studies on collocations often also aim to organize combinatory information according to semantic content with the aim of facilitating dictionary look-ups or making generalizations on combinatory behavior.

2.3.2.1 Semantic classification in collocation dictionaries

In order to facilitate access to combinatorial information, lists of lexical items combining with the headword are commonly organized into groups according to their meaning in collocation dictionaries aimed at language learners. In commercial collocation dictionaries, such as the OCD, this is done through establishing collocate groups aimed to reflect 'semantic proximity' in an *ad hoc* manner, following the lexicographers' intuition.

For instance, the entry provided for the noun *hair* in the OCD includes a group of adjectives referring to color (*auburn, black, blond, brown, chestnut*, etc.), type (*bushy, coarse, curly, fine, flyaway*, etc.), positive qualities (*beautiful, glossy, shiny, sleek*), etc. These collocation groups can appear without a label, as in the case of the OCD, or can carry explicit semantic labels as in the LCD and the MCD (sample entries of these dictionaries are shown in 3.4.2.1.B). The fact that there do not seem to be any systematic criteria underling these semantic groupings implies that no generalizations can be made as to the semantic properties of combinations constituting the whole of the dictionary content.

Another approach to describing the meaning of collocations is the application of previously defined semantic classes. A number of authors make use of this strategy in descriptive studies, mostly through adopting categories originally defined through lexical functions (LF) within the ECL framework (see Alonso Ramos 1993; Corpas Pastor 1996; Koike 2001 for the case of Spanish). Naturally, LFs are used in dictionaries produced within the theoretical framework of the ECL proper, however, there is at least one dictionary not strictly ascribed to this framework, the BBI, which also specifies the meaning of collocation groups through using some of the categories established by LFs.

More specifically, semantic description based on LFs is used in the case of four types of lexical collocations included in the BBI. Firstly, one group of VERB+NOUN$_{\text{COMP}}$ combinations is described as containing a verb, which expresses the idea of ‘creation’ and/or ‘activation’ with respect to the noun (*reach a verdict*, *inflict a wound*); secondly, another group of VERB+NOUN$_{\text{COMP}}$ collocations includes expressions with collocate verbs whose meaning can be specified as ‘eradication’ and/or ‘nullification’ with reference to the noun constituting the base (*to lift a blockade, to revoke a license*); thirdly, NOUN$_{\text{SUBJ}}$+VERB combinations included in the dictionary are defined as containing a verb which names an action characteristic of the person or object designated by the noun (*an alarm goes off, rings, sounds*), and, fourthly, NOUN+*of*+NOUN collocations described in the dictionary are specified as referring to a unit of quantification[11] associated with the second noun (*a colony of bees, a piece of advice*)[12].

[11] As mentioned in the previous section, this category of ‘quantifier’ collocation is also found in the OCD.

[12] See Alonso Ramos (1993, 161) for correspondences that can be established between the LFs used within the ECL framework and the categories of lexical collocations presented in the BBI.

The advantage of making use of predefined semantic categories is that it is allows to provide explicit semantic information which is, at the same time, generalizable to the whole of the dictionary, since semantic classes grouping collocations found in all dictionary entries come from a single set. The generalizations regarding semantic content applied by Benson et al. (1986b; 1986c) are, however, somewhat limited. As observed by Wanner (1996, 20–22), in the case of VERB+NOUN$_{\text{COMP}}$ collocations, the authors seem to omit a large number of collocations, because they do not correspond to the meaning of either of the two groups established. These combinations, however, could be included in the dictionary through introducing further semantic classes. Besides, semantic description is only present in the case of the four above mentioned groups of lexical collocations, such that the remaining types of lexical collocations are defined merely on the basis of their syntactic pattern.

In contrast with this partial classification offered by the BBI, Jousse's (2010) work involves an attempt to create a comprehensive semantic typology to be implemented in a collocation dictionary. As it is explained in more detail in 3.4.2.1.F, Jousse's aim is to enhance search possibilities in an electronic lexical database, through allowing the user to search for collocates used to express a given meaning when combined with a specific term. In order for this, she developed a semantic typology on the basis of the lexical combinatorial information represented in *DiCO* (Polguère 2000), a lexical database created within the ECL framework. Together with LFs, *DiCO* also features natural language glosses to describe the meaning of combinations; the latter are used by Jousse to identify semantic components which are then reorganized to establish semantic classes comprising her typology.

As compared to the cases of semantic classification presented above, Jousse's typology constitutes a novel approach in that it groups collocations in a purely semantic ground without attending to their syntactic pattern. This means that, for instance, the collocations *entablar amistad* 'to start a friendship' and *la amistad nace* 'friendship is born' are both included in the semantic class corresponding to the general meaning 'start'. Table 4 provides a summary of the categories included in Jousse's typology. Note that the full typology, just like *DiCO*, is not limited to syntagmatic combinations, i.e. collocations, but also covers paradigmatic lexical relations, such as synonyms, antonyms, etc. Categories applicable solely to the latter types of expressions are omitted here.

Class	Subclass		Examples
Qualifiers	**judgment**	positive	*montre exacte* ‘accurate watch’
		negative	*piste vague* ‘*unclear lead*’
	intensity	high	*attiser la peur* ‘instill fear in sb’
		low	*alléger le fardeau* ‘to ease the burden’
	other qualifiers referring to physical appearance (size, shape, color, material, etc.)		*une jupe au genoux* ‘knee-length skirt’
Phase/Aspect	**preparation**		*comploter un assassinat* ‘plot a murder’
	beginning		*une maladie se développer* ‘an illness develops’
	middle		*le cœur d’un débat* ‘center of debate’
	end		*sortir du abîme* ‘to conquer despair’
	result		*tirage* → *copie* ‘print→copy’
	reiteration		*redevenir un ami* ‘to become friends again, to rekindle a friendship’
	duration		*applaudissements prolongués* ‘prolonged applause’
Manner	**instrument/means/manner**		*arme du crime* ‘murder weapon’
	means of using or expressing ~		*au couteau* ‘with knife’, *avec gratitude* ‘with gratitude’
	ways in which ~ takes place		*admirer secrètement* ‘secretly admire’
Cause	--		*crime passionnel* ‘crime of passion’
Localization	**the place where ~ is typically exercised**		*avocat* → *cabinet, cour* ‘lawyer→office, court’; *chant* → *conservatoire* ‘singing→conservatory’
	spatial and temporal localization (≈phase/aspect)		*à bord de/dans un bateau* ‘on board of/on a ship’
Action/Event	**creation/emergence**		*lancement de un bateau* ‘launching of a ship’
	manifestation		*frémir, palpiter, trempler d`émoi* ‘to express, display, tremble with emotion’; *les pleurs coulent, jaillisent* ‘the tears run down, fall’
	use/realization/typical functioning		*consulter; utilizer les services de un avocat* ‘to consult with, take advice from a lawyer’; *la abeille bourdonne* ‘bee buzzes’
	growth or improvement		*attiser crainte* ‘to fuel fear’
	decrease or deterioration		*alléger un fardeau* ‘to lighten a burden’
	destruction/cease of functioning		*sécher les pleurs* ‘to dry up the tears’; *tomber la berrière* ‘to break down a barrier’
	non-operation of an entity, non-realization of an event, prohibition or denial		*chômer la usine* ‘to close down a factory’; *refuser, rejecter un ultimatum* ‘to refuse, reject an ultimatum’, *manquer de tact* ‘to lack tact’
	attempt		*disputer la victoria à* ‘to fight for victory’

Table 4 Semantic typology of collocations, adapted from Jousse (2010, 188)

2.3.2.2 Studies on co-occurrence patterns

While, as it has been shown up to this point, semantic classification of collocations is often seen as a means to facilitate access to combinations in a collocation dictionary, the interest of a number of authors lies in studying co-occurrence patterns in more detail, in order to establish generalizations through correlating the combinatorial behavior of lexical items with their meaning.

One such attempt involves Bosque's (2001; 2011) work, culminating in the Spanish combinatorial dictionary *Redes* (Bosque 2004b), already mentioned in 2.2.2 and 2.2.4.7. We have seen that this author defines collocations in function of whether bases co-occurring with a given collocate lend themselves to be organized into lexical classes on the basis of well-defined semantic criteria. According to the author, lexical classes themselves allow both for characterizing the meaning or meanings of bases and describing the combinatory behavior of collocates in detail. For instance, the collocate adjective *luminoso* 'illuminated, brilliant' is described in *Redes* as co-occurring with lexical classes containing NOUNS REFERRING TO OBJECTS DESIGNED FOR COMMUNICATING A MESSAGE (*señal* 'sign', *panel* 'panel', *cartel* 'poster', *baliza* 'beacon', etc.), NOUNS DENOTING CHARACTERISTICS OF PERSONS, ESPECIALLY IF THESE ARE SOCIALLY ACCLAIMED (*personalidad* 'personality', *bondad* 'kindness', *sencillez* 'naturalness', *elegancia* 'elegance', etc.) and NOUNS DENOTING A TOOL, SYSTEM OR METHOD (*estrategia* 'strategy', *procedimiento* 'procedure', *estructura* 'structure', *logística* 'logistics', etc.), among others. At the same time, the noun *leer* 'read' is described as a VERB RELATED TO LINGUISTIC EXPRESSION when combining with *en voz alta* 'loud', *de carrerilla* 'fluently', *atropelladamente* 'not fluently', as a VERB OF PERCEPTION, when it combines with *de refilón* 'with half an eye', *entre líneas* 'between the lines', *por encima* 'superficially', *de cerca* 'closely' and as A VERB EXPRESSING CONSUMPTION when it co-occurs with *ávidamente* 'greedily', *compulsivamente* 'compulsively', *con fruición* 'with joy', *febrilmente* 'feverishly' and *vorazmente* 'devouringly' (Bosque 2004a, CXLIII).

It should be noted that, similarly to the case of major commercial collocation dictionaries, lexical classes in *Redes* are established in an *ad hoc* manner through taking into account the common properties of bases co-occurring with a given predicate. Nevertheless, Bosque (2004a, CXXII) remarks that some lexical classes tend to coincide with what he refers to as "natural" or "traditional semantic classes", such as verbs referring to perception, movement, thinking or influence, or nouns denoting emotions, places or persons. This emphasizes the foundation of the author's research agenda, constituted by the idea that speakers' collocational knowledge consists of structured lists representing paradigms of bases, implying that lexical classes in a way constitute an attempt to model this knowledge.

While Bosque aims to describe collocational patterns through establishing classes of bases a given collocate co-occurs with, other authors – who conceive of the directionality of restricted lexical selection in an opposite manner (see 2.2.4.4) – attempt to correlate the semantic features of the base with its own combinatorial behavior. For instance, Mel'čuk and Wanner (1996, 209–210) claim that "lexemes with common restricted lexical co-occurrence frequently share semantic features", consequently "it must be possible, at least to some useful extent, to generalize restricted lexical co-occurrence instantiations along semantic lines". Such generalizations may, according to the authors, allow the simplification and systematization of the description of collocations in dictionaries.

These two authors coin the term *co-occurrence inheritance* for the phenomenon studied, and argue that, although grouping together all lexemes which co-occur with the same lexical items would result in non-natural classes, i.e. classes with no semantic or syntactic justification, it is indeed possible to find co-relation between shared co-occurrence patterns and semantic features (Mel'čuk and Wanner 1996, 210–211). This is exemplified through an experiment in which the co-occurrence patterns of 40 German emotion nouns, classified according to a set of semantic features, are examined. As a result, the authors are able to conclude that certain clusters including nouns with common semantic features, indeed share verbal collocates, although genuinely powerful generalizations, contrary to what is suggested by Bosque (2004b; 2011), are not possible. While, for instance, most – but not all – German emotion nouns examined can be combined with collocate verbs *empfinden* 'perceive' and *fülen* 'feel', the verb *machen* 'make', which combines with *angst* 'fear' to express the meaning 'cause' cannot be used with nearly synonymous *furcht* 'fright' or *panic* 'panic'. Similar analyses were carried out in the case of Spanish emotion nouns by Sanromán Vilas (2003) and of Hungarian support verb constructions by Vincze (2005).

Goossens (2005) and Tutin et al. (2006) take a slightly different approach when they aim to derive a typology of French emotion nouns based on shared lexical co-occurrence patterns. These authors first identify common meanings expressed by lexical combinations these nouns appear in, and then determine classes of emotion nouns on the basis of shared typical combination types – not collocates. For instance, one class includes nouns having a semantic actant expressing cause, i.e. nouns which can occur in combinations expressing causation, such as *colère* ‘anger’, *honte* ‘shame’, *dègoût* ‘disgust’, *horreur* ‘horror (disgust), *gêne* ‘embarrassment’ and *inquiétude* ‘anxiety’. The meanings conveyed by these nouns are characterized as momentary or punctual, and, according to the authors, they can also co-occur with verbs expressing control of the emotion, physical manifestation or verbal expression (Tutin 2013, 46).

The typology of collocations underlying the studies described in Goossens (2005) and Tutin et al. (2006) is further developed in the framework of the multilingual EmoBase project (Goossens et al. 2013)[13] to incorporate eight main classes: 1) INTENSITY, 2) ASPECT, 3) CAUSATION, 4) MANIFESTATION, 5) CONTROL, 6) VERBALIZATION, 7) POLARITY and 8) EXPERIENCE. Within each of these, combinations are described and grouped into subclasses on the basis of a set of semantic features. For instance, in the case of the class referring to CAUSATION, collocations are described according to features referring to *aspect* and *intensity.* When it comes to aspect, a combination can make reference to the phase of an action, i.e. the beginning, the continuation or the end, it may describe a momentary or a non-momentary and an iterative or non-iterative action. As for intensity, a collocation may refer to a high (increasing) or low (decreasing) intensity emotion.

Through the combination of these features, the authors establish the following subclasses to classify collocations under the CAUSATION dimension: a) neutral aspect causation (e.g. *to cause amazement*), b) inchoative causation, referring to the causation of the beginning or emergence of an emotion, c) collocations making reference to the causation of an increase in the intensity of an emotion, thus possessing the features of phase and high intensity, d) combinations expressing the causation of a decrease in the intensity of the emotion, i.e. having the features of phase and low intensity, and finally, e) collocations making reference to the causation of the end or disappearance of an emotion, thus having the terminative phase feature. Table 5 shows a summary of the collocation typology described in Goossens et al. (2013). In order to make more apparent the interpretation of subclasses identified within each major semantic dimension, I indicate the semantic features taken into account, together with the possible values they may take in each subclass.

[13] http://persan.rom.uni-koeln.de/emolex/emoBase/

Semantic dimension	Subclass	Example
INTENSITY applied features: *--high* *--low*	high intensity	*apreciar bastante* ‘quite appreciate’, *decepción enorme* ‘enormous disappointment’
	low intensity	*ligeramente despectivo* ‘slightly derogatory’, *levemente desconcertado* ‘slightly bewildered’
ASPECT applied features: *--punctual/non-punctual* *--phase: inchoative/gradual increase/gradual decrease/terminative* *--iterative/non-iterative* *--intensity:high/low*	punctual+ non-iterative	*momentáneamente desconcertado* ‘momentarily puzzled’, *arranque de celos* ‘fit of jealousy’
	punctual+ iterative	*volver a decepcionar* ‘to disappoint again’, *otra sorpresa* ‘another surprise’
	non-punctual	*frustración constante* ‘constant frustration’, *admiración perdurable* ‘enduring admiration’
	inchoative	*ponerse furioso* ‘to get furious’, *la frustración invade a alguien* ‘frustration invades somebody’
	phasal+ high intensity	*la cólera sube* ‘anger rises’, *desencanto creciente* ‘increasing disillusionment’
	phasal+low intensity	*la ira amaina* ‘anger subsides’
	terminative	*perder el respeto* ‘to lose respect’
CAUSATION applied features *--aspect:* *-----punctual/non-punctual* *-----phase: inchoative/gradual increase/gradual decrease/terminative* *-----iterative/non-iterative* *--intensity:high/low*	neutral	*producir asombro* ‘to produce astonishment’, *atraer la ira de alguien* ‘to attract the wrath of someone’
	inchoative	*desencadenar la ira* ‘to unchain fury’ *inspirar respeto* ‘to inspire respect’
	phasal+high intensity	*aumentar el desconcierto* ‘to increase confusion’ *fomenter el respeto* ‘to promote respect’
	phasal+low intensity	*apaciguar la rabia* ‘to appease someone’s anger’, *calmar los celos* ‘to soothe jealousy’
	terminative	*arrebatarle la alegría a alguien* ‘to ruin one’s enjoyment’
MANIFESTATION applied features: *--physical: voluntary/involuntary* *--verbal* *--external (observation of an emotion experienced by sb)*	voluntary	*gesto de desánimo* ‘sign of depression’, *manifestar su desepción* ‘to display disappointment’
	involuntary	*lágrimas de decepción* ‘tears of disappointment’, *paralizado de asombro* ‘paralyzed with astonishment’
	verbal	*grito de asombro* ‘cry of astonishment’, *suspiro de decepción* ‘sigh of disappointment’
	external	*parecer decepcionado* ‘look disappointed’, *desprecio manifiesto* ‘clear disdain’
CONTROL applied features: *--(control of) emotion* *--(control of)manifestation*	emotion	*iritación irreprimible* ‘irrepressible irritation’, *desahogar su rabia* ‘to vent one’s fury’
	manifestation	*ocultar su decepción* ‘to hide one’s disappointment’, *admiración secreta* ‘secret admiration’
VERBALIZATION applied features: *--emotive (not necessarily voluntary)* *--communicative (intentional)*	emotive	*gritar su cólera* ‘to scream one’s anger’
	communicative	*confesarse decepcionado* ‘to confess one’s disappointment’, *indignado reproche* ‘indignant rebuke’
POLARITY applied features: *--positive/negative* *--internal/external (evaluation)*	positive+internal	*sorpresa agradable* ‘nice surprise’, *sorpresa maravillosa* ‘wonderful surprise’
	negative+internal	*sorprender ingratemante* to surprise unpleasantly’, *doloroso desengaño* ‘painful disappointment’
	positive+external	*envidia sana* ‘healthy envy’, *justa irritación* ‘rightful irritation’
	negative+external	*celos injustificados* ‘unjustified jealousy’, *desprecio imperdonable* ‘unforgivable contempt’
EXPERIENCE applied features: *--presence/absence* *--neutral/additional semantic content*	presence+neutral	*sentir envidia* ‘feel envy’, *experimentar un sobresalto* ‘experience a shock’
	presence+additional semantic content	*admiración unánime* unanimous admiration’, *compartir su alegría* ‘to share someone’s happiness’
	absence	*faltar al respeto* ‘to fail to show respect’
	absence +additional semantic content	*casi despectivo* ‘almost contemptuous’, *rayar en el desprecio* ‘to border on contempt’

Table 5 Semantic dimensions of collocations containing lexical items relative to emotions according to Goossens et al. (2003)

2.3.3 Classification of collocations in the Explanatory and Combinatorial Lexicology

From the discussion offered in the previous section it is apparent that – with the exception of Jousse's (2010) proposal – the attempts at establishing semantic typologies are limited in their scope, in that they only apply to combinations relevant in a given semantic field, whereas more comprehensive lists of collocations included in collocation dictionaries tend to be organized in *ad hoc* groups, not allowing for generalizations. In contrast, as it was suggested before, LFs (Mel'čuk 1996; 1998; 2015) proposed within the ECL framework offer a formal means to describe both the syntactic and semantic characteristics of collocations in a comprehensive manner.

A LF encodes the relationship between two lexical units, the keyword (X) and the value (Y), similarly to a mathematical function: f(X)=Y. When it comes to representing a syntagmatic relation, i.e. a collocation, the base is given as the keyword of the LF, which provides the collocate or a list of possible collocates as its value(s). For instance, the LF Magn, expressing the abstract meaning of intensification, provides different values when applied to the lexical units *rain, friendship* and *dangerous* as shown in (2).

(2) Magn(*rain*) = *heavily, cats and dogs*
Magn(*friendship*) = *close, deep, great*
Magn(*dangerous*) = *extremely, highly, terribly*

There exists a list of LFs conventionally used by ECL practitioners, each of which can be specified for its general abstract meaning and a syntactic pattern depending on the part of speech of the base. These are called *simple standard* LFs, and they can be combined into *complex* or *compound* LFs; for a detailed description of these see e.g. Mel'čuk (1996b) and Alonso Ramos (1993). Consequently, LFs themselves, whether simple, complex or compound, constitute a synthetic way of describing both syntactic and semantic properties of collocations. For instance, in the examples shown in (2), we can observe that Magn does not only specify the abstract meaning of the combinations as 'intensification', but it also allows to describe the corresponding syntactic pattern. When the LF Magn is applied to a keyword being a verb (*rain*) or an adjective (*dangerous*), it prototypically gives a value which is an adverb (*heavily, extremely*, etc.), in contrast, when the keyword is a noun (*friendship*), its value is prototypically an adjective (*close, deep,* *great*). In other words, the LF Magn encodes intensifier combinations corresponding to the syntactic patterns MODIFIER+NOUN, VERB+ADVERB, and MODIFIER+ADJECTIVE.

Verbal LFs provide even more specific syntactic information. For instance, Real$_i$ takes a noun as its keyword, giving as its value fulfillment verbs, with the approximate meaning of ‘do with X what one is supposed to do with it’ or ‘fulfill (the requirement of) X’ which take the keyword as object and as subject the *i*th semantic actant of the keyword. Another verbal LF, Fact$_i$ gives as its value fulfillment verbs which take the keyword as their subject and its *i*th actant as object, see (3). Thus the combination *to prove an accusation*, where the subject of the verb represents the person making the accusation, can be described using the LF Real$_1$ considering that the actantial structure of the noun *accusation* is *X’s accusation of Y being guilty of something*; and the collocation *deny an accusation* is encoded by the LF Real$_2$, since the subject of the verb *deny* is the accused person, that is, the second actant of the noun. The noun *dream,* shown in the last two examples, has the actantial structure *X’s dream of doing or achieving Y,* consequently, the collocation *to realize one’s dream* is described by the LF Real$_1$ since the subject is the person having the dream, and *a dream comes true* by Fact$_0$, where 0 is used in the index to indicate that no semantic actant is syntactically realized.

(3) Real$_1$(*accusation*) = [*to*] *prove* [ART ~]
Real$_2$(*accusation*) = [*to*] *deny* [ART ~]
Real$_1$(*dream*) = *realize*
Fact$_0$(*dream*) = *comes true*

The principles of lexicographic description formulated within the ECL framework, including the representation of lexical relations via LFs, have been implemented in a number of lexicographic works. The four volumes of the *Dictionnaire explicatif et combinatoire du français contemporain* (Mel’čuk et al. 1984; Mel’čuk et al. 1988; Mel’čuk et al. 1992; Mel’čuk et al. 1999), as well as the *Explanatory combinatorial dictionary of modern Russian* (ECD, Zholkovsky and Mel’čuk 1984) constitute pioneering examples of such lexicographic investigation, while the *DiCo* (Polguère 2000; Jousse and Polguère 2005), DiCE (Alonso Ramos 2004), a series of multilingual terminological dictionaries including *DicoInfo* (L’Homme 2009) or *DiCoEnviro* (L’Homme and Laneville 2009), as well as the lexicographic project *Dire autrement* (Milićević and Hamel

2007), and the RELIEF Project (Lux-Pogodalla and Polguère 2011) implement the ECD format in an electronic database which can be consulted through an online interface.

While the system of LFs itself, as we have seen, represents a comprehensive and complex typology of collocations, a number of authors aiming at a better understanding of this system have proposed different classifications and typologies of the LFs themselves (see e.g. Alonso Ramos 1993; Alonso Ramos and Tutin 1996; Grimes 1990; Jousse 2010; Steele and Meyer 1990; Zholkovsky and Mel'čuk 1970). These studies will not be discussed here in more detail; an overview of them can be found in Alonso Ramos and Tutin (1996, 148–151) and Jousse (2010, 156–162), while the semantic typology based on the system of LFs and proposed by Jousse (2010) was already presented in 2.3.2.1. Typologies of LFs dealing with the semantic properties of collocations tend to group collocations which have similar meanings but are described by different LFs together in intermediate classes. This is thought to be necessary, since the ECL approach allows a rather fine-grained description of collocations, especially through the use of complex and compound LFs, which needs to be systematized for semantic studies.

As noted by Jousse (2010), from the point of view of the application of LFs in lexicography, subtle semantic distinctions of collocations are not necessarily convenient, especially considering dictionary look-ups. Instead of necessarily conceiving of each simple LF and each possible combination of LFs as constituting individual semantic classes, which could result in a typology of unmanageable size, it may be useful to group collocations in a smaller set of more generic categories, such that combinations expressing complex meanings can belong to various classes at the same time. For instance, in DiCE, the Spanish collocation *estrechar una amistad* lit. 'tighten a friendship' is represented by the complex LF CausPredPlus, whose approximate meaning is specified via the semantic gloss 'cause the friendship to be stronger'. Alternatively, this combination can be classified as belonging to a number of more generic semantic classes, such as POSITIVE QUALIFICATION, CAUSATION, GROWTH, HIGH INTENSITY or IMPROVEMENT.

In addition, through the use of generic semantic classes, it is possible to include combinations which have identical or similar meanings but different syntactic patterns in the same group(s). Since LFs represent semantic and syntactic information simultaneously, in DiCE, the collocations *coger miedo* lit. 'catch fear' and *entrar(le) miedo* lit. 'fear enters (sb)' are encoded by two different complex LFs, IncepOper_1 and IncepFunc_1, respectively, although they both make reference to a situation when 'someone begins to experiment fear'. Following Jousse's (2010) proposal, both combinations can be included in the semantic class BEGINNING. The author argues that this type of classification allows to enhance the flexibility of search options in an electronic dictionary. Jousse's (2010) proposal on enhancing dictionary look-ups as compared to more traditional collocation dictionary layouts is discussed in more detail in 3.4.2.1.F., while its implementation is the case of a learning tool focusing on Spanish collocations is proposed in 6.2.1.1.

2.4 Summary

This chapter examined the concept of collocation, tracing its evolution beginning from the emergence of the term to refer to the linguistic phenomenon of lexical co-occurrence, through the points of view of the two main descriptive traditions, the frequency-based approach and the phraseological approach. An in-depth discussion focused on comparing the views of different authors regarding the main features of collocations mentioned in the literature to the definition provided within the ECL framework, adopted in the present thesis. As it was explained, from the perspective of this theoretical framework, the key notion for defining collocations is that of lexical restriction. Collocations are conceived of as lexically restricted binary combinations, where the semantically and lexically autonomous base determines the selection of the collocate in order to express a given meaning.

The second part of the chapter focused on two aspects of the description of collocations considered to be especially relevant from a pedagogical point of view. Both the description of collocational syntactic patterns and that of the meaning of combinations are essential when it comes to presenting combinatorial information in collocation dictionaries. Accordingly, I reviewed combinatorial dictionaries and the work of a number of authors dealing with these aspects. Finally, I introduced the system of LFs, used to describe both the syntactic and semantic properties of lexical combinations within the ECL framework, and discussed some of the implications of its implementation in dictionaries. The issue of the description and classification of collocations will be revisited in 3.4.2.1.B in relation to the structure of collocation dictionaries.

Chapter 3. Collocations in second language learning

3.1 Introduction

The aim of this chapter is to look at collocations in the context of second langue (L2) learning. It begins with an overview of the main theoretical considerations highlighting the importance of learning multiword expressions – among them collocations – in a L2 context, and goes on to examine the results of studies dealing with L2 learners' collocation competence offered in the literature. Following this, it provides a brief review of pedagogical approaches to teaching collocations, and, finally, concludes with a description of existing resources supporting collocation learning and use.

3.2 The role of multiword expressions in language: L1 acquisition, L2 learning and language use

Perhaps one of the greatest challenges when it comes to reviewing and interpreting the literature dealing with collocations and multiword expressions in language acquisition pertains to terminology – hence the title of the present section. A major difficulty results from the generalized use of the term *formulaic language* in the literature. Following Wray (2002, 9), *formulas* or *formulaic sequences* are commonly defined as multiword expressions which are prefabricated, i.e. stored and retrieved from the memory as a whole, facilitating language processing. Consequently, the term represents a broad category, which is often understood to include collocations as a subtype, while some other terms referring to types of formulaic expressions, such as *idiom*, *cliché*, *phraseme*, etc. are also in use (Wray 2002, 9). Thus, it may concluded, that much of the research on formulaic language is relevant for the case of collocations, and therefore, it is included in the present literature review. Another problem is caused by the varying definitions of collocations themselves used by researchers. Given that authors may rely on different notions of collocation, closer to either the frequency, the phraseological or a mixed approach makes the comparison of results of empirical studies especially problematic (see e.g. Henriksen

2013, 44). Finally, it should be also noted that, regardless of the terminology used, studies often fail to provide a precise enough description of the exact type of expressions studied.

As noted by Durrant (2008, 38–39), although multiword expressions or formulas have been on the second language teaching agenda for a rather long time, they have often been considered as merely a transitory feature of novice learner language (see also Wray 2000, 463; Wray 2002, 191–192). It has been assumed that the memorization of multiword strings serves as a temporary tool that allows instant communication, before chunks are replaced by genuine creative language use (see e.g. Krashen and Scarcella 1978). This view is based on the idea that learning, resulting in creative language use, should entail the mastery of single vocabulary items and grammar rules; consequently, the memorization of multiword sequences is not considered a useful long-term language learning strategy.

A contrasting view holds that multiword expressions have a central role in language. The growing interest in teaching and learning multiword sequences resulted in an increasing number of studies, as well as in giving a more prominent place to these expressions in language curriculums and teaching materials. Durrant (2008:40) identifies three main motivations for teaching multiword expressions, including collocations, in the literature (see also Nation 2001, 318–325). The first of these concerns the role of multiword units in language acquisition and language learning, more precisely, theories of language acquisition claiming that non-fragmented multiword sequences or chunks play an important role in the acquisition of – at least – the first language (e.g. Ellis 1996; Wray 2002; Tomasello 2003). The second and third rationales focus on the role of multiword units in language use in general. The second argument emphasizes the fact that multiwords appear to make up an important portion of language, thus, sufficient knowledge of formulaic sequences is essential in order to achieve native-like production. The third rationale is related to findings on language processing which point in the direction that familiarity with multiword units facilitates fluent language production. The following subsections explore theoretical considerations and research findings related to the role of collocations in language acquisition, learning and use from the perspective of these three arguments.

3.2.1 Mutiwords in the language acquisition and language learning process

The first of the motivations identified by Durrant (2008:40) for teaching formulaic sequences to language learners is the key role attributed to them when it comes to acquiring a linguistic system, at least in the case of the first language. A number of authors (e.g. Tomasello 2003; Lieven and Tomasello 2008; Bannard and Lieven 2012; N. C. Ellis 2012) claim that it is through the memorization and subsequent analysis of multiword sequences that abstract patterns of language enabling creative language use are learnt. This view, often labeled as the *usage-based approach*, stands in contrast with nativist theories, which propose the existence of innate abstract grammatical knowledge guiding L1 acquisition. Instead, it is hypothesized that while no innate grammatical categories whatsoever exist, children acquire a repertoire of strings of language from which they induce increasingly abstract patterns through form and meaning overlaps identified in these. Thus grammatical regularities, and, consequently, creative language use emerge from the abstract patterns identified in the memorized strings.

Evidence from research supporting the usage-based theory – for reviews see Tomasello (2003), Bannard and Lieven (2012) and Ellis (2012), among others – resulted in the reassessment of the importance of teaching formulaic language and memorized strings to second language learners. If the process of L2 learning follows a similar pattern to that of L1 acquisition, formulas should not be seen as a mere shortcut to achieve effective communication by the non-proficient learner, but in fact constitute the key to the acquisition of the linguistic system. Accordingly, in his EFL syllabus, Willis (1990, ii–iii) suggests that effective teaching methods should take advantage of learners' ability to work out the grammar when exposed to the most common patterns of language. In a similar vein, Nattinger and DeCarrico (1992, 27) claim that grammar is at least to some extent "acquired through generalizing, and learning the restrictions on the generalization" from formulaic sequences.

Wray (2000, 470–471), however, calls attention to the fact that, while teaching syllabuses tend to assume learners' ability to make generalizations about grammar on the basis of formulaic input, results of empirical research sometimes point in the opposite direction (cf. Granger 1998, 157). Yorio (1989, 68), for instance, concludes on the basis of corpus data that adult L2 learners are not able to use formulaic language to further their grammatical development. On the contrary, others, such as Myles et al. (1999) claim to have found evidence for L2 learners using and subsequently breaking down complex chunks in the process of acquiring abstract grammatical structure; for a review of further similar studies see e.g. Durrant (2008, 53) and Ellis (2012, 31–34). In sum, the somewhat contradictory results of empirical studies suggest that one should be cautious when it comes to establishing a correspondence between the processes involved in L1 and L2 acquisition.

Wray (2000, 471–472; 2002, 144–145) suggests that this inconclusive nature of research outcomes is owing to the fact that learners are often treated as a homogeneous group, disregarding their age or whether they are exposed to the foreign language in a naturalistic or a classroom context. This is problematic since different learner groups can be characterized by very different learning patterns, with young children's learning process being more similar to L1 learning than that of adults. As for the case of adult learners, Wray (2002, 144) suggests that while formulaic sequences do not contribute to the learning of grammatical patterns in the case of naturalistic L2 learners, classroom-taught adults are able to break down formulaic strings. However, as she captures in her model of "the creation of the lexicon in classroom-taught L2" (pp. 208-210), the purpose of this analysis is to access the lexical constituents of the expression, which are then stored separately, while the information concerning the fact that these lexical items form a sequence and the links between them are not retained. Consequently, Wray's conclusion is that formulaic sequences do not contribute directly to the acquisition of grammatical rules in the case of adult L2 learners.

In sum, it seems more plausible that multiword expressions do not play the same role in L2 learning as in L1 acquisition, when it comes to the acquisition of grammar (see also Durrant 2008, 51). Ellis (2012, 31–34) and Wray (2002, 205–206) provide a number of possible explanations supporting this idea. The most relevant is perhaps the one pointing out the fact that L2 learners are not learning from scratch about abstract grammatical categories or linguistic structures, since they have already acquired a linguistic system. This implies that learners can establish correspondences between elements and constructions in their L1 and the L2 being learnt, therefore, they may be more likely to attempt creative language use based on the already familiar patterns. Additionally, it is argued that, adult learners over literacy age tend to see word-sized units as more salient and more manageable than complex strings.

In view of the fact that there is no conclusive evidence as to whether formulas aid the development of the learners' language system – in a way similar to what is assumed in first language acquisition –, Durrant (2008, 57) warns that researchers must be cautious about referring to this acquisition route as one of the major arguments for teaching formulaic language. While focus on teaching formulas "may provide learners with a useful mental phrasebook of utterances for specific situations”, it may also “leave them unable to adapt their language to new situations or to express more novel ideas”.

3.2.2 The role of multiword expressions in language use

The second and third rationales highlighting the importance of multiword expressions mentioned by Durrant (2008, 40) concern not the learning process, but language use in general. One concerns to the claim that formulaic sequences constitute a considerable part of any discourse, while the other refers to their positive affect on fluency. These two aspects are in fact parallel to the two linguistic capacities Pawley and Syder (1983) describe as *nativelike selection* and *nativelike fluency*.

In the idiom principle, already mentioned in 2.2.1, Sinclair (1987, 320–321) articulated the idea that language cannot be properly described by open-choice models, since speakers often make use of “semi-preconstructed” phrases. Pawley and Syder refer to the same phenomenon (1983, 192–199) as *nativelike selection* pointing out that a large portion of language is in fact idiomatic, in the sense that otherwise grammatical sentences often cannot be said to be native-like, i.e. natural-sounding. For instance, the sentence shown in (5), which constitutes the paraphrase of (4), does not sound natural. This implies that, language learners need to “learn a means for knowing which of the well-formed sentences are nativelike” (p. 194). In fact, as it is shown in 3.3.2.1, one of the findings of corpus studies is that non-native language largely differs from native language with respect to the use of multiword expressions.

(4) I’m so glad you could bring Harry.

(5) That you could bring Harry gladdens me so.

What further emphasizes the importance of formulas is their wide-spread nature in language. Although quantitative data offered in different corpus studies aiming to establish the percentage of multiword expressions in discourse varies considerably, there seems to be sound evidence for the frequent use of formulaic language. For instance, Biber et al. (1999, 993) found that 30% of words appear in lexical bundles, i.e. recurrent expressions, in a conversation corpus, while 21% in a corpus of academic prose. Analyzing a considerably smaller set of data, Erman and Warren (2000, 37) found that an average of 55% of words occur in a prefab, i.e. a conventionalized expression, in their corpus comprised of oral and written texts. Focusing on the type of combinations that concern the object of this thesis, Cowie (1991; 1992) reports that 37.5% to 44% of all verb plus direct object combinations in page one news stories, and 46% in feature articles can be classified as restricted collocations or idioms, while Howarth (1996, 122) finds that frequent verbs appear in restricted collocations in 36% of all occurrences in native academic prose.

The arguments that learning formulaic language contributes to more native-like language use, and that formulas constitute a considerable percentage of language easily lead to the conclusion that the more formulas are used by language learners, the more natural their output will be. This claim, however, is rather over simplistic, as it is not necessarily supported by empirical evidence. While, Kaszubski's (2000, 242) data show a correlation between language proficiency and the amount of restricted collocations used by learners, Nesselhauf (2005, 234–236) does not find such correlation neither when taking into consideration the number of years of L2 instruction nor the time spend abroad by learners. Furthermore, as we will see in Section 3.3.2.1, over-reliance on certain formulas has in fact been found to constitute a characteristic of non-native language use.

As mentioned above, the second aspect of language observed by Pawley and Syder (1983, 205) is *nativelike fluency*. These authors argue that the explanation to how native speakers can produce long stretches of discourse without hesitation lies in that certain multiword sequences are stored as single memory units in the long term memory (see also Ellis 1996). Therefore, they hypothesize that formulaic language, aside from being a wide-spread phenomenon, also constitutes an important benefit, contributing to fluent language production. Evidence for this is provided, for instance, by the commonly cited work of Kuiper (1996; 2004). This author found that auctioneers and sports commentators, whose professions require them to produce fluent speech under time pressure, use a large amount of formulaic expressions in their professional discourse.

Conklin and Schmitt (2012) review empirical studies providing evidence for formulaic chunks being processed differently from non-formulaic language. Importantly, research has found evidence for the facilitating effect of frequent sequences both in perception and language production, and in the case of different types of multiword expressions, including collocations. Authors generally claim, similarly to Pawley and Syder's (1983) original hypothesis, that the facilitating effect of common multiword sequences can be explained by the fact that they are holistically represented in the mental lexicon. The same idea is reflected in, Wray's (2002, 132–135, 207) model of the creation of the lexicon, which predicts in the case of native speakers that formulaic sequences are holistically stored and are broken down during the language acquisition process only when and to the extent necessary. As for the particular case of collocations, Wray (2002, 211) claims that her model provides an explanation for why these expressions are problematic for learners. Assuming that native speakers tend to store multiword sequences as single units, analyzing them only as necessary, collocations in the native lexicon can be conceived of as fully formulaic sequences, which become loosened during the language acquisition process. In contrast, adult language learners' collocations are to be seen rather as single-word items which become paired when a collocation is learnt.

Conklin and Schmitt (2012: 55-56) observe that reasons other than holistic memory storage have also been suggested to explain processing advantages found in the case of formulaic sequences. One of these refers to the predictability of certain frequent sequences –, which constitutes the phenomenon exploited by probabilistic language processing models. Yet another proposal is that of Hoey (2005), according to whom formulaicity should be explained in terms of strength of association and priming effects between the components of the expressions.

Results of studies (e.g. Siyanova and Schmitt 2008; Schmitt et al. 2004) suggest that the above mentioned processing advantages are not only observable in the case of native speakers, but also in proficient language learners, although, to a lesser extent. At the same time, less proficient learners have been found to process formulaic sequences in a word-by-word manner, similarly to non-formulaic language (for a review of related empirical research see Conklin and Schmitt 2012).

As we have seen, Wray's (2002, 206–210) model of "the creation of the lexicon in classroom-taught L2" predicts the lack of holistic storage in language learners, suggesting that they tend to analyze formulaic sequences straight away, instead of storing them as single units. According to this author some of the factors which may contribute to word-size units being the salient target in the foreign language learning process are the different communicative needs of adult language learners as compared to those of young children, the focus on form approach in language learning, and the post-literacy age of learners (Wray 2002, 205-206).

Nevertheless, the fact that non-native speakers have been found to process at least some multiword sequences with an ease similar to what is observed in the case of native speakers, suggests that learners may store certain formulas holistically, while they analyze and process others as creative language. This raises the question of what causes the difference in the storage of different expressions. In fact it might be the case that such differences cannot only be found in the case of language learners, but also in the language processing of native speakers. In an experimental study Schmitt et al. (2004, 138) observe that, not only is there a difference in the probability with which different sequences are stored as single units in the case of native speakers, but individual speakers themselves also differ in which expressions they seem to store holistically. This suggests that each speakers' mental lexicon may contain a unique inventory of formulaic sequences.

This lack of homogeneity in how different multiword sequences are processed and stored is often explained by correlating processing advantages with the frequency of occurrence of a given string in the input. Siyanova and Schmitt (2008, 445–446), for instance found that native speakers processed high frequency verb+noun collocations significantly faster than mid-frequency combinations. Similarly, Durrant (2008) concludes in a word association study that higher frequency of occurrence in corpus can reliably predict the mental storage of high frequency collocations. For a review of further studies correlating frequency and processing advantages in multiword sequences see Conklin and Schmitt (2012, 50–52).

Results of other studies, however, point to a more complex relation between frequency and memory entrenchment. Ellis et al. (2008, 389–391), for instance, finds that, in the case of native speakers, it is not raw frequency, but association strength represented by mutual information score that best correlates with speed of processing of formulaic expressions, which explains why lower-frequency but strongly associated sequences also present processing advantages. The authors argue that this is so, because the effect of frequency reaches a ceiling once a native speaker is exposed to sufficient amount of language input. On the contrary, in the case of non-natives, highly associated low-frequency expressions are not established in the memory, simply because they are not encountered sufficient times in the considerably lower amount of input these speakers are exposed to.

To sum up, if the knowledge of formulaic sequences contributes to more fluent language production, and frequency in the input raises the probability of holistic storage, it may be possible to improve non-native fluency through the increased exposure to formulas. Although there exist pedagogical proposals for enhancing fluency through the memorization of multiword sequences in communicative drills (e.g. Gatbonton and Segalowitz 2005), as claimed by Taguchi (2008, 135), empirical evidence is so far too scarce to prove their effectiveness. This author describes a study which shows that specific instruction focusing on “grammatical chunks”, i.e. grammatical constructions containing open slots, increases the fluency of elementary level students of Japanese. In a case study Wood (2009) describes the evolution of the oral production of a L1 Japanese student of English, participating in a specific fluency workshop, designed to promote the noticing and memorization of formulaic sequences. Although the findings of such a small-scale study cannot be considered conclusive, the results show an increase in the use of formulaic language and the rate of fluency in the students’ narrative, both of which can be likely attributed to the fluency workshop.

3.2.3 Summary: Why teach formulaic language?

This section has observed three main motivations for teaching formulaic language commonly discussed in the literature. Some authors argue that formulas constitute a central component in the acquisition of the linguistic system; furthermore, there is a rather strong consensus concerning the importance of these expressions for attaining native-like language production and native-like fluency. Although not all of these three claims are strongly supported, taken together, they make a good case for a greater emphasis on formulas in the foreign language classroom.

As it has been noted, the object of study of the present thesis, collocations, constitutes a mere subset of the group of expressions referred to as formulaic sequences, which are characterized in Wray’s (2002, 9) definition as holistically stored in the speakers’ memory or lexicon. It should be noted in this respect that, while Wray (2002, 211) herself does conceive of collocations as holistically stored expressions, there is no clear agreement in the literature regarding this aspect, and consequently, regarding the extent to which collocations can be considered to overlap with formulas (see Henriksen 2013, 41). Regardless, it seems rather justified that at least the rationales relating to native-like production and fluency are highly relevant in the case of collocations as understood in the present thesis. In fact the same motivations are given by researchers studying specifically collocation phenomena when arguing for the importance of collocation competence in the case of foreign language learners (cf. Durrant 2008; Granger 1998; Nesselhauf 2005; Siyanova and Schmitt 2008).

3.3 Studies on language learners' knowledge and use of collocations

As it was discussed in 2.2.4.7, authors ascribed to different descriptive frameworks of the phraseological approach describe the main difficulty posed by collocations to language learners as resulting from lexical restrictedness, which affects language production. The previous sections provided a somewhat different insight, from the perspective of language acquisition theory and psycholinguistic research on language processing, where collocations are often considered as a subtype of formulaic expressions. Here, the main reasons given for learners' potentially different processing and mental representation of multiwords are the insufficient exposure or low frequency of occurrence of these expressions in the input, and L2 learners' tendency to focus on committing individual words, instead of complex units, to memory.

Language learners' apparent focus on individual words – which may also be partially teaching-induced – implies that in L2 production collocations are potentially constructed from individual items, and, given the restricted nature of these expressions, may not necessarily be nativelike. Further issues relevant in the case of collocation learning mentioned by e.g. Henriksen (2013, 40–42) include the relative transparence of these expressions, their lack of salience in the input and learners' lack of awareness of the existence of restricted combinations. These three aspects are naturally related: since collocations do not cause a major comprehension problem provided their component lexical units are known to the learner, they are not likely to be noticed, in addition, since learners' are not aware of the phenomenon of collocations, they do not look for them in the input.

Gyllstad (2007, 50) notes that despite the fact that the problematic nature of collocations in second language learning is often emphasized in the literature, the number of empirical studies investigating learners' knowledge of collocations is not especially abundant. It should be added that studies focusing on learners of languages other than English are notably scarce. In general, collocation knowledge and collocation use are investigated via two main approaches: studies using testing methodology and learner corpus studies, the former concentrating on receptive and/or productive knowledge, the latter obtaining collocation production data from naturalistic learner texts.

In what follows, I review studies using tests in order to evaluate collocation knowledge, as well as corpus studies on learner's collocation use. In doing so, I dedicate special attention to factors affecting language learners' collocation production and the analysis of collocation errors, since these aspects are particularly relevant to the empirical studies described in the following chapters of the present thesis.

3.3.1 Studies testing collocation knowledge

This section provides a brief overview of the results of studies testing language learners' knowledge of collocations. The studies mentioned here were chosen according to whether the type of targeted expressions corresponds to the focus of the present thesis. Thus, for instance, Gitsaki's (1996) work will not be considered, since her most relevant findings concern grammatical collocations, as defined in Benson et al. (1986b, see 1.3.1), while results relevant to lexical collocations only cannot be straightforwardly extracted. Keshavasz and Salimi's (2007) study represents a similar case; while the majority of expressions (e.g. *Fine Arts, alarm clock, safety belt, radio set*) included in the test presented by Hussein (1990) are considered as compounds or quasi-idioms in the theoretical framework followed here. Other than the types of expressions learners are being tested for, when considering the results of these studies it is also important to bear in mind the fact that testing methodology has become more rigorous over the last decades. As noted by Moreno Jaén (2009, 260), while studies carried out in the 1990s can be considered as the pioneers of investigating collocation competence, research methodology has received more emphasis in the 2000s, with studies using a thorough piloting phase for test development, and tests which themselves contain a higher number of carefully chosen items are administered to a greater number of participants to enhance reliability and validity.

Evaluating collocation competence through testing allows for approaching learner knowledge from two different angles: reception or production. Receptive knowledge is usually measured through multiple choice tests or decision making tasks where participants need to judge the correctness or idiomaticity of given combinations. These are also commonly referred to as recognition tests, since they ultimately measure participants' ability to recognize correct or familiar collocations. Tests assessing productive knowledge usually use translation tasks and cloze tests where participants are required to supply a member of a collocation – usually the collocate – or the whole combination. Certain authors concentrate solely on investigating either receptive (Granger 1998; Mochizuki 2002; Gyllstad 2007; Siyanova and Schmitt 2008; e.g. Eyckmans 2009) or productive knowledge (e.g. Bahns and Eldaw 1993; Biskup 1992; Farghal and Obiedat 1995; Revier

2009; Nizonkiza 2012), while others test both types of competence (e.g. Bonk 2000; Koya 2005; Moreno Jaén 2009).

One aspect of collocation competence tackled by a number of researchers is the comparison between learners' and native speakers' ability to discriminate between more idiomatic or frequent combinations and less prototypical or incorrect ones. Granger (1998, 152) asked L1 French learners of English and native speakers to choose the acceptable collocates of 11 amplifiers from a list of 15 adjectives, and found that combinations indicated to be especially salient in the native data were marked by considerably fewer learners, e.g. *readily available* was marked by 43 out of 56 native participants while only by 8 out of 56 learners. As a result, the author concluded that learners have a weak sense of salience, hence they are not able to judge what constitute significant combinations. In a similar study presented by Siyanova and Schmitt (2008), native and non-native participants were asked to judge the commonness of collocations. It was found that while non-native speakers did rate frequent collocations on the higher half of the scale than non-frequent ones, the mean scores given differed less than in the case of native speakers. Furthermore, native speakers were found to be able to reliably discriminate between high frequency and mid-frequency collocations, whereas non-native speakers were not. Thus the authors concluded that their results indicate an emerging collocation knowledge in the case of learners, which is not as accurate as that of native speakers. Data obtained also suggested that the length of naturalistic L2 exposure correlates with learners' ability to rate collocations for commonness.

Another matter pursued by several authors is the relationship between collocation competence and general language proficiency. To this end some researchers compare the test performance of student groups belonging to different study levels. Gyllstad (2007), for instance, found a correlation between study level and test performance in a large scale study focusing on the receptive knowledge of English verb+noun collocations of upper-intermediate and advanced level Swedish students. The author hypothesized that the explanation for this correlation lies in that in the case of low-proficiency learners links between L2 lexical forms and conceptual representation are predominantly mediated via L1 translation equivalents, thus members of collocations tend to be processed separately, while the role of this type of mediation decreases in more proficient learners (Gyllstad 2007, 245–250). Correlation between study level and collocation competence was also found in another large scale study by Moreno Jaén (2009), investigating both receptive and productive knowledge of L1 Spanish EFL students, and in Revier (2009), a study which introduced a novel test format designed to measure whole collocation production and was administered to L1 Danish learners of English. A number of authors assessed general proficiency through administering an external test, instead of study level. Both Bonk (2000) and Nizonkiza (2012) found correlation between learners' performance on a collocation production test and their scores on the TOEFL exam. Gyllstad (2007) carried out a follow up study to his large scale experiment, using a limited number of subjects, and found that results obtained from his collocation recognition test correlated with learners' scores on the reading component of the Certificate of Advanced English exam. Finally, some authors studied the evolution of collocation competence in longitudinal studies. Eyckmans (2009) observed an improvement in L1 Danish EFL learners' ability to discriminate between idiomatic and non-idiomatic combinations after 60 hours of instruction; similar results were obtained by Mochizuki (2002) in the case of L1 Japanese students who were tested for receptive collocation knowledge at the beginning and at the end of an academic year, during which they received about 75 hours of EFL instruction.

The evolution of collocational competence was not only compared to general proficiency, but also to vocabulary size. Mochizuki (2002), for instance, asked participants in his longitudinal study to complete a vocabulary size test, especially developed for Japanese EFL learners. While the author found that learners' receptive collocational competence had significantly improved over the academic year, only a non-significant increase was observable in their vocabulary size. Therefore, he concluded that collocational knowledge – as part of word-knowledge – changes more easily than meaning knowledge (Mochizuki 2002, 128). Koya (2005), who assessed both receptive and productive collocational competence of his L1 Japanese subjects, found both constructs to correlate with vocabulary size, measured by the same test developed by Mochizuki (2002). In a similar vein, Gyllstad (2007) found significant correlation between receptive collocation knowledge and vocabulary size, as well as vocabulary depth, measured through a modified version of the Vocabulary Levels Test (Nation 2001; Schmitt 2001) and Word Association Test (Read 1993), respectively. The author concluded that the existence of a strong relationship between vocabulary size and receptive collocation knowledge is not surprising since single words constitute the building blocks of collocations, while subjects with larger vocabulary size can also be assumed to have had a greater amount of L2 exposure (Gyllstad 2007, 239).

A few authors examined collocations produced in elicitation tasks more closely to gain an insight into learners' production strategies. Biskup (1992), for instance, observed important differences between L1 German and L1 Polish EFL learners' performance. On the one hand, she found that Polish learners refrained more often from providing answers in the translation task while Germans often used paraphrases or "descriptive answers", such as *make the clock working* instead of *winding a watch*. Consequently, she concluded that, possibly due to different instructional styles, the Polish learners in her study can be characterized as low risk-takers, who aim at more accurate language. Besides, the author also observed that, probably as a result of the perceived distance between Polish and English, L1 Polish learners relied more on semantic transfer, while German students produced more transfer errors resulting from similarity (e.g. **crunch/crunk nuts* instead of *crack nuts*). Biskup also highlighted that Polish learners produced more transfer errors in general, which she attributed to the lack of more creative production strategies observable in the case of German students. In my opinion, this finding can also be explained by the different degrees of similarity between the language pairs, since it is possible that L1 transfer from German results in correct collocations more often than transfer from Polish. Similar strategies to those manifested by Biskup's Polish students were found by Koya (2005) in the case of L1 Japanese EFL students, who also appeared to be low risk-takers and relied on L1 transfer to a lesser extent than learners in other studies, again probably due to the perceived distance between the target language and their L1.

Bahns and Eldaw (1993) administered a test measuring productive knowledge of verb+noun collocations through a cloze test and a translation task to L1 German advanced-level students of English. These authors focused on the use of paraphrases instead of restricted lexical combinations as a production strategy, establishing that correct renderings of collocations in a translation task were provided by a paraphrase in less than half of all correct answers. They found that some collocations (*serve a sentence, withdraw money, refuse admission, take a call, pay compliments*) were less amenable to paraphrasing than others (*keep a diary, cancel an order, reject a proposal, do damage, whip cream, achieve perfection*). Farghal and Obiedat (1995) identified a number of production strategies in a fill-in-the-blank exercise and a translation task administered to L1 Arabic EFL students. The most frequently used one of these was described as the use of a synonymous lexical element (e.g. **steady color, *stable color, *static color* instead of *fast color*), which according to the authors indicates learners' lack of awareness of collocational restrictions. It was observed that both paraphrase and L1 transfer occurred more often in the case of the translation task – clearly due to task-effect. It must be noted, however, that in certain cases (e.g. **firm color*), the distinction of transfer from the use of synonymous expressions does not seem straightforward.

Lastly, test results are also evaluated in order to establish what types of collocations pose more difficulty for learners. Some of the factors taken into account are syntactic pattern, semantic transparence, word frequency and L1 congruence. Firstly, while most studies use as test items collocations corresponding to one or a limited number of syntactic patterns – typically verb+noun –, Moreno Jaén (2009) establishes a difficulty scale taking into account different syntactic patterns. This author observed that her subjects had most difficulty with noun+noun combinations, followed by noun+verb and adjective+noun collocations, while they provided the highest proportion of correct answers in the case of verb+noun combinations. It should be noted, however, that certain syntactic patterns were rather underrepresented in the study, therefore, the significance of these results is questionable. Secondly, Mochizuki (2002), Koya (2005) and Revier (2009) all found learners to perform better with transparent collocations than non-transparent ones in their studies focusing on reception. Thirdly, Nizonkiza (2012), who used carefully selected collocations as test items, chosen according to the frequency of their lexical elements, found that learners had less difficultly with items belonging to more frequent word levels. It is not clear, however, to what extent these findings are relevant for collocation knowledge, since participants only had to supply one member of the combination in the production test, therefore, it may be the case that the frequency of the single element and not that of the collocation is relevant for test performance. Finally, both Farghal and Obiedat (1995) and Koya (2005) reported a lower error rate in the case of L1 congruent combinations, a finding that, as we will see, is supported by learner corpus studies.

3.3.2 Learner corpus studies focusing on collocation use

Learner corpora are defined by Granger (2002, 7) as “electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose”. Considering especially that, as we have seen in the previous section, the object of SLA research is often constituted by a limited amount of data obtained from a competence test, these collections of larger quantities of learner language constitute a highly valuable resource and contribute to obtaining detailed descriptions and hence a better understanding of learner language. Most available learner corpora – and, as we will see in this section – research using learner corpora, are of L2 English, and concern written language. I have knowledge of only a handful of very recent studies – besides the one whose results are presented in the following chapter – dealing with collocations in Spanish learner corpora; these are Orol González and Alonso Ramos (2013), Pérez Serrano (2014) and Uriel Domínguez (2014). For an overview of available corpora of L2 Spanish see Mendikoetxea (2014).

Learner corpora and methods used in corpus linguistics are especially suitable for studying collocation use. In studies investigating this aspect of learner language, authors present essentially two types of data. Firstly, learners' collocation use is studied on a quantitative basis, and is compared to native speakers' production to discover patterns of under- and overuse, as well as differences in the typicality of collocations used, in addition, a few authors also study the relationship between learners' proficiency level and the amount of collocations produced. Secondly, learners' collocation use is described in terms of correctness, on occasions through a more detailed error analysis. These two approaches present in corpus studies are explored in more detail in the following sections.

Note that, as in the case of competence tests, studies concerned with multiword expressions not fitting into the definition of collocation applied in the present thesis will not be considered in this review. Such is the case of Gitsaki (1996), whose study concerns both lexical and grammatical collocations, nevertheless, her most compelling findings concern only the latter, and Yuldasev et al. (2013), using a Spanish learner corpus collected from L2 Spanish speakers in the US to study the use of multiword units, especially discourse markers.

3.3.2.1 Quantitative differences between non-native and native collocation use

The initial hypothesis behind corpus studies focusing on collocation use from a quantitative perspective is usually derived from the assumption that language learners' production would be more dominated by the Sinclairian (1987: 320–321) open choice principle than that of native speakers' who make use of "semi-preconstructed" phrases more often (see e.g. Granger 1998: 146; Durrant 2008: 165–166). Consequently, language learners are expected to use a lower number of collocations than natives. However, results of studies concerning mostly EFL learners reveal a more complex picture of learners' use of different types of combinations.

An early study inspired by Sinclair's claims is Yorio (1989). After having examined three different sets of data from learners with different L1 backgrounds, the author claims that contrary to the initial assumption that language learners do not acquire formulaic language, he did find an extensive use of what he calls "conventionalized expressions". He also emphasizes, however, that these expressions found in the learner texts are not error free. Consequently, he speculates that chunks are not learnt as wholes by non-native speakers, but they are analyzed from the beginning. When comparing the writing of Spanish L1 immigrant students to that of Spanish L1 students living in their native Argentina, Yorio (1989, 67–68) finds that the second group produced more grammatically accurate language and used more idioms and, in general, a more native-like language. This seems to highlight the importance of explicit instruction in order to achieve accuracy in foreign language production. According to the author, idiomaticity, i.e. native-like production is closely related to the occurrence of a higher amount of collocations, which he defines as habitual syntagmatic combinations, in the learner texts.

As it was discussed in 3.2.2, the use of semi-preconstructed phrases or formulas is not only hypothesized to contribute to native-like production, but also to fluency. Learner corpus studies generally do not deal with this latter connection. As an exception, Nesselhauf (2005), in a study on the use of verb+noun combinations in texts produced by L1 German learners of English, observed that learners did not use more collocations in essays written under time pressure, i.e. they did not increase their use of collocations to enhance fluency, and, concluded that learners did not appear to have much automatic control of collocations (p. 230). Nesselhauf's (2005) corpus based observations thus do not seem to lend support to the correlation between fluency and the amount of multiword expressions used.

The hypothesis that learners make use of less formulaic language, in particular, collocations, than native speakers has been further investigated in studies – which unlike Yorio (1989), whose claims are rather impressionistic in this respect – contrast learner corpus data with comparable texts produced by native writers. Underuse of collocations was observed in a number of studies by different authors, most of whom, however, concentrated on a specific type of combination, usually corresponding to a given syntactic pattern. Granger (1998), who can be considered as the pioneer of using modern corpus linguistic methodology for studying language learners' collocation use, focused on amplifying adverbs ending in *–ly* used as modifiers (e.g. *perfectly natural, closely linked, deeply in love*), and found that these are in general significantly underused in the writing of L1 French learners, as compared to texts composed by native speakers. Similar findings are reported by Howarth (1996) in the case of restricted verb+noun combinations. This author found that non-native writers underused restricted collocations as compared to natives (24% vs. 35%[14]), while they used a higher proportion of free combinations corresponding to the same syntactic pattern (67% vs. 60%). Altenberg and Granger (2001) found that L1 French and L1 Swedish learners used *make* as a support verb considerably less often than native English speakers; this underuse being especially prominent in the case of nouns referring to 'speech' and 'verbal communication', e.g. *argument, claim, point, statement, assumption,* which are in fact among the most frequent nouns to combine with *make* in the native texts. Finally, Laufer and Waldman (2011) found less than half as many verb+noun collocations in essays collocated from Israeli learners of English than in the native reference corpus (5.9 vs. 10.2%).

Notably, in a study using data from the *Corpus Escrito del Español L2* (CEDEL2, Lozano 2009; Lozano and Mendikoetxea 2013) corpus comprised of essays written by L1 English learners of Spanish, Orol González and Alonso Ramos (2013) report having encountered only a negligible difference in the amount of collocations used as compared to native essays. This finding is especially interesting in the light of that, in contrast with previously mentioned studies which only take into account collocations corresponding to one particular syntactic pattern, here all collocations were extracted from the corpora. Note, however, that these authors do not offer data with regard to the distribution of collocations grouped according to syntactic patterns. In a similar vein, Siyanova and Schmitt (2008) report that they found no significant difference in the amount of adjective+noun collocations used by L1 Russian students and their native English speaker counterparts.

While, as we have seen, underuse of collocations by non-native speakers is attested at least in the case of particular types of combinations, the opposite phenomenon, overuse has also been observed in specific cases. Granger (1998), for instance, reports that combinations containing adverbs *completely* and *totally* are significantly overused in learner texts. The author suggests that the overuse phenomenon can be explained by the

[14] In one instance, the amount provided is 36% (see Howarth 1996, 157).

fact that the expressions involved function as “safe bets”, since they display few collocational restrictions both in English and in the learners’ L1. In contrast, the significant underuse of the adverb *highly* is explained by the fact that its literal L1 equivalent is much less frequent. In sum, over- and underuse of specific expressions seem to point towards that most collocations used by learners are congruent with L1 collocations. Lorenz (1999), who studied adjective intensification in argumentative essays written by L1 German advanced students of English concentrating primarily on stylistic features, found that learners used a higher overall amount of adjective intensification than natives. He claims that this overuse can be accounted for as a stylistic characteristic specific to German learners, resulting from their tendency to information over charge.

Importantly, Lorenz (1999) also found his subjects to rely on “a limited number of high frequency stock items”. This is evidenced by their overall repertoire of collocations, measured by type-token ratio being lower than that of natives, who command more different intensifier types. Kaszubski (2000, 243), who studied the use of six English high frequency verbs (*be, do, have, make, take* and *give*) in L1 Polish students’ writing, found that his subjects tended to overuse simple vocabulary in general, as well as combinations made up of vocabulary items and combinations which are congruent with L1 forms. Finally, Nesselhauf (2005), although without contrasting her data with native corpora, reports her impression that learners appear to overuse frequent collocations, such as e.g. *solve a problem, have time, have a chance,* and, based on Hasselgren’s (1994) concept of *lexical teddy bears*, she coins the term *collocational teddy bears* for this overuse phenomenon (Nesselhauf 2005, 68–69).

Besides observing differences between non-native and native collocation use, similarly to test-based studies, certain corpus studies also aim at investigating whether collocation competence evolves parallel to general language proficiency. Howarth (1998b) compares the results of a proficiency test and his learner corpus data, but fails to find a correlation between proficiency level and the amount of conventional language used. In a very tentative manner, he suggests that his results can be explained either by the failure of the test to establish learners’ proficiency level accurately or by that the use of conventional language may in fact develop separately from proficiency, given it is an “individual matter of style” (p.36). Kaszubski (2000), in contrast, claims to have found such correlation, observing that the use of free combinations decreases with the increase of proficiency/expertise in writing (p. 242), while the overall use of restricted collocations increases (p. 244). Similarly, Laufer and Waldman (2011), found Israeli advanced level EFL learners to use more collocations (6.2%) than their intermediate (5.3%) and basic level peers (4.3%).

Learner corpus research presented up to this point, in general terms, focuses on comparing the amount of collocations used by native and non-native speakers, or that used by groups of non-native speakers corresponding to different proficiency levels. Another point of comparison refers to the frequency and/or idiomaticity of collocations. Kaszubski’s (2000) observation that learners show a preference towards using collocations comprised of high frequency lexical items has already been mentioned above. Orol González and Alonso Ramos (2013), report similar findings; their data showing that the mean frequency of lexical items constituting the bases of collocations used by L2 Spanish speakers is lower than that of bases in the native corpus. In his study, Lorenz (1999) compared adjective+intensifier combinations used by learners and natives in terms of mutual information (MI) scores. As mentioned in 3.2.2, MI can be taken as a measure of idiomaticity, since higher MI scores characterize infrequent but strongly-associated pairs typically used by native speakers. Accordingly, the author found that the average MI score for combinations in the non-native corpora (MI=7.41) was 20% lower than in the native corpora (MI=9.22) examined. Finally, Durrant (2008, 165–183) compared the use of directly adjacent premodifier-noun pairs (adjective+noun and noun+noun combinations) in non-native and native texts, by using raw frequency, as well as association measures such as t-score and MI score. He found that native writers made use of a higher percentage of low frequency combinations, i.e. combinations that were found to occur less than 5 times in the BNC, than non-native writers. According to the author, this can be interpreted as non-natives being more conservative in using collocations in the sense that they are less likely to coin combinations than natives, preferring to rely on pairings attested in the input. Non-native writers were also found to make at least as much use of frequent combinations, corresponding to the higher t-score bands, as natives, while they underused collocations with high MI scores. From this, Durrant concluded that what characterizes learners collocation use “is not that [their] language is missing high frequency phrases, but rather that it makes too little use of these lower-frequency but strongly-associated items” (p. 183). Therefore, he suggested that the learners’ perceived problems with collocations are not due to their incapacity of acquiring word combinations – as suggested by Wray and other authors (see 3.2.2) – instead, it can be simply put down to the lack of sufficient input. In other words, collocations not encountered sufficiently are not acquired, regardless of how their salience to native speakers.

3.3.2.2 Learner collocation errors

We have seen that researchers examining the nature of collocations produced by learners have observed a number of common traits. These are the use of combinations containing higher frequency lexical items, the repetitive use of certain high frequency combinations, referred to as "collocational teddy bears" by Nesselhauf (2005, 88–89) and the preference for using combinations which are congruent with the learners' L1. Another aspect that has often been studied in relation to learners' collocation production is the amount of erroneous combinations, and the nature of collocation errors.

Nesselhauf (2005) and Laufer and Waldman (2011) found that about one third of collocations in their corpora were erroneous, whereas a somewhat lower error rate is reported by Uriel Domínguez (2014), who studied both grammatical and lexical collocations in essays collected from students of Spanish at B2 proficiency level. This author identified 196 correct and 32 erroneous collocations in the texts of L1 French students, while 242 correct and 31 erroneous collocation instances were found in L1 Dutch students' essays. A considerably higher error rate – 144 correct collocations as opposed to 31 "improvable" and 153 incorrect combinations – was found by Serrano Pérez (2014) in the writing of A1-A2 level Spanish students with English as L1. Naturally, error counts are highly dependent on both the types of collocations and the types of errors (lexical and/or grammatical) taken into account in a given study, in addition to learners' proficiency level. Empirical data is rather scarce with regard to this last aspect, since most studies deal with a relatively homogeneous group of learners. Laufer and Waldman (2011, 665)), whose corpus contained essays from basic, intermediate and advanced learners of English found certain correlation between proficiency level and the amount of erroneous collocations produced.

Only a few authors provide a detailed description of the nature of collocation errors found in corpora, however, even those who do not, usually point at L1 influence being a main factor in producing erroneous combinations. Laufer and Waldman (2011, 665)) claim that L1 influence appears in about half of the collocations identified in the corpus, in the case of all three proficiency levels studied. Martelli (2006) also observes that many of the collocation errors identified in her L1 Italian learner corpus can be attributed to L1 influence (e.g. *nowaday situation* from *situaziona odierna*; *animated discussion* from *discussione animata*, *give successful results* instead of *obtain* from *dare dei buoni risultati* etc.)

Moreover, authors find most recurrent errors in the corpus to be transfer induced. Nevertheless, L1 influence, i.e. transfer does not always result in the production of erroneous combinations. As mentioned earlier, a number of authors in fact observed the overuse of combinations congruent with learners' native language. Granger (1998, 150–151), for instance, establishes that "stereotyped combinations" found in learner texts typically have a direct translation equivalent in L1 or they are "lexically congruent" (see also Kaszubski 2000, 243). In fact, learners have also been found to rely to a greater or lesser extent on L1 transfer depending on the perceived distance of their L1 and the L2. Kaszubski (2000, 246), as we have seen, who studied essays produced by L1 Polish, Spanish and French learners of English, claims that from his data it appears that Polish students, whose L1 is more distant from English were more reluctant to coin new combinations, hence they produced less erroneous combinations, but overused familiar collocations to a greater extant. Similarly Uriel Domínguez (2014, 49), who found a higher number of errors in the production of L1 French learners of Spanish than that of L1 Dutch learners, claims that these results might be explained by the similarity of the two romance languages.

As it was mentioned, some authors go into more detail when describing collocation errors. Since one of the main features of the corpus study presented in the following chapter is the classification of learner errors according to a typology specific to collocation errors, the remainder of the present section will be dedicated to an overview of descriptions provided by other authors. As we will see, most of these have several points in common with the error typology which will be presented in Chapter 4.

Yorio (1989) submitted one of his corpora containing essays of non-native students residing in the US to a detailed error analysis (pp. 63-64). He enumerated a number of error types characteristic of "conventionalized expressions": 1) grammatical errors, e.g. **take advantages of, *he had chance*; 2) lexical choice errors, such as **made a great job, *make a great influence, *put more attention to*; 3) mixed idioms, as in **give up their freedom of mobility* instead of *give up their mobility* or *freedom of movement*, **it always strikes the mind of the employer* instead of *to strike sb* or *to cross the mind of sb;* 4) phrases used with wrong meaning, e.g. *in this way* with meaning 'for this reason', *in* *addition to* with meaning ‘in order to’; 5) “attempted idioms” as in the case of *at the end of the road* to express ‘ultimately’, **turn on his mind as a blank* instead of *their minds go blank*; and, finally, 6) missed register errors, as in **a better society without crime and headaches*.

Howarth (1996, 145–146), whose study focuses exclusively on verb+noun collocations, on the contrary to Yorio (1989), considers that grammatical errors, such as the cases of missing prepositions in e.g. **interfere* [*with*] *the process, *compensate* [*for*] *the lack, *respond* [*to*] *students' need*, are not especially relevant for the study of collocations. However, he identifies errors resulting from the combination requiring a certain morphological form, e.g. *losing his *tie with his own culture group,* claiming that it is in fact difficult to decide whether these should be classified as grammatical of lexical errors. In other borderline cases the error results from the extension of an established collocation to a derivative of one of its elements, e.g. **draw a conclusive comment* from *draw a conclusion.*

Eventually, limiting his analysis to cases of lexical errors, Howarth (1996, 146–156), distinguishes five main error types. 1) The first of these is described as “clearly nonce forms which fail to communicate a clear meaning as a result of the unnaturalness of the lexical co-occurrence”, such as e.g. **accomplish interest, *collect interest.* 2) The next error type consists of the confusion of delexical or support verbs, as e.g. in **do attempts* and **do a measurement.* 3) The third type is constituted by the use of an “unconventional” (rather than a clearly erroneous) collocation in which the verb is used in a figurative sense: *?break the limitations, ?pursue factors, ?raise factors.* 4) The fourth type concerns overlapping collocations, that is, cases when the erroneous verb and the target verb are semantically too dissimilar to be easily confused (**attach a role* instead of *give a role*), however the erroneous verb can express the target meaning with a different but semantically similar noun (**attach a value*), which, in addition, shares at least one collocate verb with the target noun in the given context (*attach/assign a value, assign/give a role*). Similar cases are **cause an effect* and **perform a project* which would be erroneous combinations in the overlapping clusters *cause/produce a reaction//produce/have an effect* and *carry out/perform a task//carry out project*, respectively. 5) Finally, the last type of lexical collocation errors identified by Howarth is blending, i.e. instances where the collocate verb is confused in the case of two semantically close nouns, whereas there is no verb that could collocate with both nouns.

Therefore, the erroneous combination **achieve tasks* is a blend produced through confusing the combinations *achieve goals* and *perform tasks,* and **draw a correlation* results from the confusion of *draw a comparison* and *make a correlation*. These last two types of errors are also described in Howarth (1998b, 37).

Although they provide a relatively detailed description of different types of errors identified in the corpus, the authors mentioned above do not consider explicitly which part of the combination was affected by a particular type error. In this sense, Lorenz's (1999) approach to describing the errors affecting intensifier+adjective combinations introduces a new dimension of classification. This author draws an explicit distinction between errors affecting the adjective – the base of the collocation, according to the terminology applied in this thesis – and the intensifier – the collocate. The error types affecting adjectives (pp. 134-141) are the following: 1) morphological errors, such as in **very controverse*[15] instead of *controversial*, **quite unlogic* instead of *quite unlogical*, 2) "mis-hyphenations" or "malformed compounds", as e.g. **really suspiciously-looking* instead of *suspicious-looking,* 3) incorrect uses of participle forms as adjectives, as in **very influenced, *totally organized*, 4) semantic errors, i.e. uses of existing but erroneous adjectives to convey a given meaning, such as in (*becomes) *more and more actual* meaning *topical, current.* The author claims that most of these errors result from "creative lexical innovation", and constitute ad-hoc coinages which are often constituted through L1-L2 transfer. The error types affecting intensifiers are the following (pp. 142-162): 1) word formation or lexical usage errors, as in e.g. **detailly organized *purely red, *widely spread*, 2) position errors, i.e. cases of erroneous word order with adverbs *enough, just, so, quite, rather, too*, 3) semantic intensifier errors, where erroneous adverbs are chosen to express intensification, e.g. in the case of **apparently bored newscaster* instead of *evidently bored*, 4) erroneous choice or misuse a marked intensifier in an innovative expression: *computers [...] are *efficiently stupid, this kind of news shows are *boringly stiff*; finally, the last type of error is described as 5) "incongruous grading", referring to the use of maximizer adverbs with adjectives requiring grading intensifiers as in **not quite exciting, *absolutely easy/silly/stupid, totally damaged*, or boosters modifying non-gradable adjectives, as in

[15] Note that Lorenz's (1999) analysis concerns all occurrences of intensifier+adjective combinations, some of which would not be considered as restricted lexical combinations, i.e. collocations following the theoretical framework used in this thesis.

**extremely huge, *very horrible/delicious, really impossible/necessary* or *extremely different.*

Among the studies reviewed here, it is Nesselhauf (2005) who provides the most detailed description of collocation errors. This author, focusing exclusively on verb+noun combinations characterizes errors from two different perspectives: the element of the combination affected by the error and the nature or origin of the erroneous element used. The author finds that the element most frequently affected by errors is the verb, followed by the noun and cases when the whole combinations is considered inappropriate – these are, in most cases, lexical errors –, while grammatical errors, such as those involving the determiner, noun complementation, preposition, etc. are less frequent (p.72). See Table 6 for some of the most common error types established by Nesselhauf to describe the erroneous element.

Erroneous element	Type of error	Examples
Verb	inappropriate single verb	**comsume drugs, *disturb phone-calls*
	inappropriate use of a phrasal verb	**open up jobs* instead of *create*, **have out a smell* instead of *give off*
	inappropriate use of a prepositional verb	**hop to a conclusion* instead of *leap to, ride on bicycles* instead of *ride*
	use of a verb judged to be superfluous in a given context	*you would not think twice about a separation because it is easy* [**to do*]
Noun	inappropriate choice of number	**pass one's judgements, have* [*less*] **chances to find a job*
	use of an inappropriate or a non-existent nominal element	*fulfil sb's *wishes* instead of *dreams, enjoy our *private atmosphere* instead of *privacy, get a *divorcion* instead of *divorce*
	superfluous uses of nominal elements	*I did that *procedure* [*getting up at night to smoke…*] *about twice or three times a night*
Determiner	Inappropriate, superfluous, missing, etc.	*go through *a higher education, *an own apartment, *run risk of*
Noun complementation		**find new ways* [*of sth*], *have a right of* instead of *to*
Preposition introducing an adverbial or a complement		**get into contact with, *put sb into prison*
Global deviations	use of a stretched verb construction instead of a corresponding verb	[*a computer*] **had a breakdown* instead of *break down*
	inappropriate verb + noun combination instead of another expression	*fall into a fit of laughter* instead of *laugh*
	inappropriate verb+noun combination instead of another verb+noun combination	*They [people who go on holiday] get new impressions* instead of *broaden one's horizons*
	structural deviations	*give a rest* instead of *give us a rest, make sb friends* instead of *make friends with sb*

Table 6: Examples of error types established by Nesselhauf (2005) to describe the erroneous element of the collocation

As for the origin of collocation errors, the author examines the nature of L2 lexical material used and the possible effect of L1 influence. She notes that single word forms not existing in L2 were introduced only in rare occasions, the only examples being *declench a war* instead of *start*, *superate hard times* instead of *go through*, *divorcion* instead of *divorce* and *dispens* instead of *disposal*. In contrast, errors resulted more frequently from the recombination of existing L2 material to from non-existent phrasal verbs, prepositional verbs or N-*of*-N compounds (p 166). Nesselhauf (2005) also observes that sometimes existing L2 lexical combinations were used inappropriately or to convey a wrong meaning (*take measures* instead of *measurements, make fun of sb* instead of *make fools of, take time* instead of *make time, name the price* instead of *tell the price, pass no judgement* instead of *make no assessment*). Consequently, she hypothesizes that learners' are often familiar with L2 combinations, but, at the same time, they lack a precise knowledge of their meaning and/or use (pp. 167-168). The author also provides quantitative data regarding the relevance of formal and semantic links in erroneous lexical choice. She finds that formal links between the erroneous and the appropriate target element are more frequent in the case of nouns (41%) than verbs (16%), while sematic links are considerably high in the case of both nouns (59%) and verbs (51%) (pp. 169-174). Similarly to Howarth (see above), Nesselhauf also examines cases of blending, although her definition of the phenomenon corresponds to a narrower category than that of the previous author. Blending, according to Nesselhauf, involves the creation of a novel expression through confusing two L2 chunks which have at least one element in common. Thus the expression **break a new record* is produced through blending *break a record* and *set a new record*; other examples are: *the title of this essay carries the meaning* (*this essay carries the title* + *carry the meaning*) and *take into consideration seriously* (*take into consideration* + *take seriously*).

Regarding transfer, Nesselhauf (2005, 181) establishes that L1 influence on the inappropriate element or on at least one of the inappropriate elements is likely in the case of 51%, while it is possible but less likely in a further 2% of all erroneous collocations. In the case of lexical collocation errors, 52% of errors in the collocate verb and 47% of errors in the noun were likely due to L1 influence. Taking into account semantic and formal links between L1 and L2 elements, overall 3 main types of L1 influence are identified: 1) the use of a verb where the translation equivalent would be appropriate in the German collocation but it is not appropriate in the English collocation, and the verbs may or may not be formally related (*put questions* from *Fragen stellen, *bring themselves in danger* from *sich in Gefahr bringen, make homework* from *Hausaufgaben Machen*), 2) the use of a verb which is not a translation equivalent of the German verb used in the equivalent combination but is formally related to it (**become problems with* from *Problem bekommen mit* 'have problems with'), and 3) the use of a verb with several translation equivalents, one of which, but not the translation equivalent used, would be appropriate in the collocation (**take over responsibility* from *übernehmen*, instead of *take on*; *follow their aims* from *verfolgen*, instead of *pursue*) (p.191).

3.3.3 Summary: Knowledge and use of collocations in a foreign language

The last sections provided a brief overview of the findings of studies assessing language learners' collocation competence and collocation use. These are mainly limited – with a few exceptions – to the case of EFL learners, and can be divided into studies using testing methodology and learner corpus studies, the former concentrating on receptive and/or productive knowledge, the latter obtaining collocation production data from naturalistic learner texts. Results of collocation competence tests show that language learners' ability to discriminate between idiomatic and non-idiomatic combinations is not as reliable as that of native speakers, while their collocation competence seems to correlate with general proficiency and vocabulary size.

Corpus studies show that, as predicted by theoretical approaches such as Sinclair's open choice and idiom principles or Wray's models of the "creation of the lexicon", learners appear to use less collocations than native speakers – although evidence is lacking as to whether this is the case with e.g. collocations corresponding to any type of syntactic pattern. Contrary to expectations, non-native speakers are also found to use certain types of combinations more extensively than natives. These are often described as favorite collocations or "teddy bears" since they are easily acquired, usually frequent combinations learners can use with certainty. As with collocation tests, general language proficiency has been found to correlate with more native-like collocation use in the case of naturalistic production, represented by higher amount of collocations used.

Erroneous responses in collocation tests – especially those focusing on productive knowledge – as well as analysis of collocation errors extracted from learner corpora provide an insight into language use strategies and specific difficulties posed by collocations. The crucial role of L1 transfer in language learners' collocation use is demonstrated both by the high amount of transfer errors and the facilitating effect observed in the case of congruent collocations. Other strategies applied by learners to convey the target meaning expressed by a restricted lexical combination are paraphrase, substitution of an element of the combination by a synonymous lexical item, blending or confusing the elements of not fully acquired expressions, etc. As we have seen, collocation errors can be especially revealing and can be described in much detail. The analysis of collocation production, focusing on collocation errors will be further pursued in the case of a Spanish learner corpus in Chapter 4.

3.4 Teaching and learning collocations

The previous two sections discussed the main benefits attributed to the learning of multiword expressions and the results of studies on language learners' knowledge and use of collocations, respectively. The present section focuses on pedagogical proposals for teaching and learning collocations, concentrating especially on resources which can be exploited for autonomous learning, such as dictionaries, language corpora and online learning tools.

3.4.1 Pedagogical approaches to teaching collocations

A number of pedagogical proposals for foreign language teaching adopted the views of e.g. Pawley and Syder (1983) and Sinclair (1987), which emphasize the role of multiword chunks in language, and challenge the traditional dichotomy of grammar and the lexicon. In the case of English as a foreign language, Willis (1990), Nattinger and DeCarrico (1992) and Lewis (1993; 1997) all propose to foreground lexis, and in particular, formulaic sequences in the teaching syllabus. What is common to the work of the three authors is their claim that lexis and grammar are not separable, and that, in fact, language learners should be trained in identifying patterns in the input, thus deriving grammatical knowledge. While Nattinger and DeCarrico (1992) concentrate on the usefulness of a type of multiword expressions with specific pragmatic functions, they denominate *lexical phrases* in teaching conversation, Willis (1990) is interested in developing material to teach the most frequent words of English, together with their characteristic patterns, which he stipulates to be the most common in the language. In order to make vast lexico-grammatical information more manageable, this latter author proposes to organize co-occurrence patterns into generalizable frames, so that, in the case of collocations, a given frame can be applied to specific groups of lexical items constituting classes, e.g. the frame DELEXICAL VERB+*a*+NOUN applies in the case of the verb *give* and the class constituted by all nouns co-occurring with it, as e.g. in *give a glance* (Willis 1999, 129–131).

Among the three above mentioned pedagogical proposals, Lewis' (1993; 1997) Lexical Approach is the one that addresses the issue of teaching collocations in most detail. This author claims that collocation teaching should constitute an important part of the language syllabus, such that single lexical items, in particular, high content nouns should be introduced together with verbs and adjectives they form strong collocations with (1993, 110). He justifies this by arguing that the procedural knowledge of a word "involves mastering its collocational range and restrictions on that range" (p. 119). In Lewis (2000a, 134–136), the author emphasizes that one difficulty of learning collocations is that while some of these expressions may appear to be logical or obvious word combinations, they may not be so from the point of view of the speaker of another language. For instance, the combination *have a baby* seems completely transparent, whereas the utterances *she has a baby* and *she is having a baby* have different meanings, and in fact have different corresponding translation equivalents in e.g. Hungarian: *kisbabája van* 'she has a baby' and *kisbabát vár* 'she expects a baby', respectively. As being able to identify and observe language phenomena is one of the main pillars of the Lexical Approach paradigm, Lewis claims that learners should be taught to single out collocations through awareness raising activities, since this helps "to obtain maximum benefits from the input to which they are exposed" (1993, 120), and he also describes language corpora and dictionaries as important resources for collocation learning (2000b). Activities and classroom experience resulting from putting the ideas formulated within the Lexical Approach with regard to teaching collocations into practice have been described in e.g. Conzett (2000), Lewis (1997, 143–146), Hill et al. (2000) and Woolard (2000).

Following the example of pedagogical development in the case of EFL, the *Plan curricular del Instituto Cervantes* (PCIC, Instituto Cervantes 2006), the official syllabus for SFL, developed and applied by the Cervantes Institute, also places an emphasis on including multiword expressions among the proposed target lexical items. Its authors underline the fact that the lexical contents of the syllabus were chosen in accordance with theoretical claims that an individual's lexicon comprises a vast number of "semi-constructed chunks", in addition to simple lexical units. This is reflected in a novel organization of meaning related lexico-grammatical knowledge, which, instead of single words, relies on concepts, leaving space for the inclusion of relevant lexical combinations. Furthermore, different combinations containing a given lexical item are incorporated in different consecutive proficiency levels, with the intention of gradually increasing the depth of knowledge of the lexical item in question. For instance, the verbal collocates of the noun *duda* 'uncertainty' are included in parts of the syllabus corresponding to different proficiency levels under the label "notions of existence: certainty, uncertainty", see Table 7.

B1	**B2**	**C1**	**C2**
tener una duda 'have a question'	*tener dudas* 'have doubts'	*plantear una duda* 'raise a question' *resolver una duda* 'clear up a doubt *sembrar dudas/la duda* 'sow doubts/the doubt'	*albergar dudas/la duda* 'harbour doubts/the doubt' *suscitar dudas/la duda* 'raise doubts/the doubt' *alimentar dudas/la duda* 'nurture doubts/the doubt' *despejar una duda* 'erase/eliminate doubt'

Table 7: Collocations of the noun *duda* 'uncertainty' corresponding to different proficiency levels included under "notions of existence: certainty, uncertainty "in PCIC

When it comes to the methodological aspects of introducing collocations in the SFL classroom, the most detailed pedagogical proposal was developed by Higueras García (2006; 2007). This author describes a five-step sequence specifically designed for teaching collocations, starting from the introduction of the concept of collocation to activities which help memorizing and prompt the use of new combinations. This sequence is accompanied by a set of eighty proposed activity-types, together with concrete examples. Further teaching sequences and exercises have been designed by e.g. Álvarez Cavanillas (2008), Fernández Lázaro (2014), Ferrando Aramo (2009), Navajas Algaba (2006) and Pacheco López (2003).

An essential component of all methodological proposals or activity sequences for presenting collocations to foreign language learners is awareness raising, such that they often contain an explicit explanation of the concept of collocation and offer example activities where learners are asked to identify collocations or collocation errors (Higueras García 2006; Woolard 2000; see e.g. Conzett 2000; Ferrando Aramo 2009). It was already mentioned above that being able to identify collocations in the input is suggested to be beneficial to learners within the Lexical Approach. Given that, as suggested by Lewis

(2000a, 134–136) as well as other authors, collocations may seem trivial or are not particularly salient, it is convenient to bring the phenomenon of arbitrary idiomatic combinations to learners' attention. Another important argument exposed e.g. by Hill (2000, 61) and Woolard (2000, 34) is that classroom time is too limited to explicitly teach enough single lexical items or, for that matter, lexical combinations, therefore, learners should be instructed in vocabulary learning techniques, such as singling out collocations in L2 discourse, which they can apply autonomously both in and outside the classroom context (see also Boers and Lindstromberg 2012, 88). In this respect, encouraging empirical results are presented by Eyckmans et al. (2007), who found that EFL learners instructed in text chunking during the period of a school year were able to identify significantly more formulaic sequences in a text than peers who had not received such instruction. The authors conclude that awareness raising in fact appears to help learners gain more appreciation of the syntagmatic dimension of language. Finally, Woolard (2000, 35) notes that collocations are well suited for autonomous learning, since given the arbitrary nature of these expressions, the role of the teacher is rather limited, apart from selecting and listing target items. He claims that since "collocation is mostly a matter of noticing and recording", once trained to do so, language learners can take their learning in their own hands.

3.4.2 Collocation learning resources

As discussed above, authors dealing with collocations within the pedagogical perspective generally highlight the importance of explicitly introducing language learners to the notion of collocation, as well as equipping them with strategies they can use to further their collocational competence through autonomous learning. It should be noted, however, that in addition to learning to autonomously identify collocations as interesting or target items in the input – which is the outcome classroom activities often aim for – language learners should also be familiarized with resources and reference tools they can exploit for collocation learning, and be instructed in their use.

Dictionaries and language corpora constitute important tools that can be used by learners in order to find information concerning the combinatorial behavior of lexical items. Learning activities dealing with the use of these two types of resources are suggested by e.g. Hill et al. (2000), Lewis (2000b) and Woolard (2000). Other researchers have aimed at creating interactive learning tools specialized in collocations (e.g. Wu, Witten, and Franken 2010; Wu, Franken, and Witten 2010; J. C. Wu et al. 2010; Potthast, Trenkmann, and Stein 2010; Wible and Tsao 2010). The following sections provide an overview of these three types of collocation learning resources, i.e. dictionaries, corpora and collocation learning tools.

3.4.2.1 Collocation dictionaries

Dictionaries constitute the default comprehensive reference tool available to the learner outside the language classroom, when it comes to vocabulary. Collocations, however, are a specific type of entity, only the subject of systematic description in a particular type of dictionary, the collocation dictionary. Nevertheless, general learner dictionaries, which are more often available and generally more familiar to language learners can also be applied for retrieving combinatorial information, even though to a limited extent (see e.g. Howarth 1996, Lewis 2000b, 200–202; 170; Woolard 2000, 36–38). That is why, although focus here is placed on the description of collocation dictionaries, other types of dictionaries are also discussed.

The following subsections consider the potential of dictionaries as collocation learning tools, starting from a brief discussion of how collocational information is introduced in different types of lexicographical products. This is followed by a detailed description of the structure and content of the lexical entries offered by specific English and Spanish language combinatory dictionaries. This allows both to compare individual dictionaries and to obtain a general idea concerning the notions involved in organizing and presenting combinatorial information in lexicography. After this, I review the results of usability studies which provide an insight into language learners' ability to manipulate different dictionaries, as well as the suitability of the presentation of collocational information. Finally, I briefly discuss Jousse's (2010) proposal, already mentioned in 2.3.2.1, aiming to enhance the collocation dictionary format through providing more dynamic access to combinatory information.

A. Collocations in different types of dictionaries

Monolingual learners' dictionaries (MLDs) do not only focus on the decoding process, through providing sufficiently accurate and transparent definitions, but also aim to give useful information for encoding. Importantly, over the last years, these dictionaries began to place more emphasis on the phrasal nature of language, providing lists of phrases and word combinations containing the headword. Collocations can be typically identified in usage examples – either highlighted or not –, they may be compiled in lists in a specific part of the lexical entry, while sometimes they are presented in a dedicated collocation box. Given that the focus here is on resources specialized on collocations, different strategies applied by MLDs for introducing collocations in lexical will not reviewed; for a detailed discussion see Dziemianko (2014, 260–264).

The main drawbacks of MLDs when it comes to describing collocations are that, due to space restrictions, they only provide a limited amount of information, and that such information is not always easy and quick to find. Therefore learners have to be trained to locate and identify target collocations. Lewis (2000b, 201), for instance, mentions that – as opposed to collocation dictionaries – monolingual dictionaries do not typically include the expressions *spare time* or *save time* under the headword *time* which constitutes the base of the collocation, but in the lexical entry of the collocate. A strategy which does not necessarily benefit encoding (see 3.4.2.1.B). Similar observations can be made about bilingual dictionaries. While they tend to include an increasing number of collocations, finding particular combinations is far from trivial. In the *Oxford Online Spanish-English dictionary*[16], we find for instance the collocation *dar cuerda a un reloj* ‘wind a watch/clock’ under the headword *reloj,* but not under *dar* or *cuerda*, while *plantear una duda* ‘raise a question’ is included in the entry for the verb *plantear* and not the noun *duda.* Lewis (2000b, 201–202) notes that the production dictionary *Longman Essential Activator* (Rundell 1997) provides a more straightforward access to combinatorial information, which, naturally is still of limited amount as compared to specific collocational dictionaries. Finally, it should be noted that electronic dictionaries, especially those equipped with more advanced search options allowing to query the content of entries and examples clearly make lookups less cumbersome in all cases.

Collocation dictionaries were already mentioned in 2.3, when discussing different ways of classifying lexical combinations. These constitute an effective and rather more comprehensive reference tool, as they are specialized in describing the particular linguistic phenomenon of restricted lexical combinations. Learners, however, often have to be explicitly introduced to this type of dictionaries, since, on the one hand, many of them may be ignorant of their existence, and, on the other, the content and organization of these lexicographical tools differs largely from that of more familiar MLDs or bilingual dictionaries. Collocation dictionaries usually do not provide definitions of headwords,

[16] http://www.oxforddictionaries.com/

whose meaning is specified, if at all, in concise glosses, similarly to the case of the combinations themselves presented.

B. The structure of the lexical entry in collocation dictionaries

The aim of this section is to discuss two main aspects concerning how combinatory information is presented in different collocation dictionaries: 1) the orientation of the description of combinations, i.e. whether collocations are listed in lexical entries corresponding to the base or the collocate, and 2) the internal structure of the lexical entry, with respect to the organization or classification of collocations. These are especially relevant from the point of view of the quality of information access, i.e. the search paths users need to follow in order to find a given combination in a dictionary.

The dictionaries examined are the three most recent English collocation dictionaries in the market, namely, *Oxford Collocations Dictionary* (OCD, McIntosh et al. 2009), *Macmillan Collocations Dictionary* (MCD, Rundell 2010), the *Longman Collocations Dictionary and Thesaurus* (LCD, Mayor 2013) and three Spanish collocation dictionaries learners can easily gain access to, *Redes. Diccionario combinatorio del español contemporáneo* (*Redes*, Bosque 2004b), *Diccionario combinatorio práctico del español contemporáneo* (*Práctico*, Bosque 2006) and *Diccionario de colocaciones del español* (DiCE, Alonso Ramos 2004). Note that, the first two Spanish dictionaries are related in that *Práctico* constitutes a "practical" version of *Redes*, therefore, it is more suited to be used by language learners, instead of concentrating on linguistic description through the classification of lexical information, as the first version does. Barrios Rodríguez (2007) provides a detailed comparison of the structure and content of the two Spanish combinatory dictionaries. For an extensive description and comparison of English collocation dictionaries see McGee (2012) and Nuccorini (2003). Buendía Castro and Faber (2014) analyze and compare three major English and Spanish collocation dictionaries, while Ferrando Aramo (2012) provides a detailed description of English, Spanish, French and Italian paper and electronic combinatory dictionaries.

The orientation of the description of collocations refers to whether dictionary entries are constituted by bases as headwords and a list of corresponding collocates or the other way around. Some authors (e.g. Ferrando Aramo 2012; Heid 2004) refer to this as lemmatization, understanding that a collocation can either be lemmatized under the base or the collocate in the dictionary. These two types of organization of collocational information are characterized in e.g. Alonso Ramos (2002, 86–93) as catering to language production or to comprehension, respectively (see also Hausmann 1979, 191–192; 1989, 1010). It is assumed that when producing a collocation, the speaker first selects the lexical item corresponding to the base, since this item is semantically autonomous, while the selection of the collocate is restricted by the base. On the contrary, when it comes to reception, the learner is expected to look up the member of the collocation, a single lexical item he or she is not familiar with. Since the meaning of the collocate can often be idiosyncratic or figurative, it is assumed that it is more helpful if the combination is found in the lexical entry of this item. Thus, if a dictionary is aimed at aiding learners' use of collocations in language production – which is generally considered to be the more problematic aspect – it should be structured accordingly. Some authors, such as e.g. Buendía Castro and Faber (2014, 207), however, emphasize that both the directions of production and reception are useful. Alonso Ramos (2002, 86–93) provides a description of what encoding and decoding collocation dictionary entries should ideally look like. The author claims that encoding collocation entries specifying the combinatorial behavior of bases should contain more detail and be organized according to the principles of the ECL framework (see 2.3.3 and 3.4.2.1.C), while decoding entries corresponding to collocates should contain less detail, being limited to lists of bases the given collocate can be combined with.

From among the above mentioned dictionaries, the OCD is most clearly a production dictionary, as it exclusively includes entries for bases, i.e. nouns, verbs and adjectives. As it was discussed in the previous chapter (see 2.3.1.1), noun entries, for instance, include verbal, nominal and adjectival collocates, as well as governing prepositions. E.g. the entry of the noun *hair* contains a list of adjectives (*auburn*, *curly*, *shiny* hair etc.), verbs (*lose*, *brush*, *comb* *one's hair;* hair *grows*, *falls* etc.), nouns functioning as quantifiers (*lock*, *wisp of* hair) as well as nouns functioning as attributes (hair *loss*), see (6). Nevertheless the electronic version, distributed in CD-ROM with the print edition, also includes concise entries for collocates with a list of possible bases. For instance, in the case of the verb *comb*, we find a list of nouns such as *hair*, *internet*, *wool* and *wreckage* which are used as direct object, see (7). While other entries, such as that of the adjective *excellent* contain both collocations where the adjective functions as the base, in this case, with verb and adverb collocates, and as a collocate, combined with a series of nouns which function as the base of the combination. Importantly, the two types of information are presented in different formats, see (8).

(6) hair *noun*

ADJ. **auburn, black, blond, brown, chestnut, dark, fair, ginger** (*BrE*) **, golden, grey/gray, grizzled, jet-black, light, raven** (*literary*)**, red, sandy, silver, silvery, white, yellow | bushy, coarse, curly, fine, flyaway** (*esp. AmE*) **frizzy, kinky** (*AmE*)**, nappy** (*AmE*)**, shaggy, spiky, straight, thick, wavy, wiry, wispy | beautiful, glossy, shiny, sleek | disheveled/disheveled, dry, dull, fuzzy, greasy, matted, messy, scruffy, tousled, unkempt, unruly, untidy** (*esp. BrE*)**, windswept** ◇ *a new shampoo for dull or dry ~* ◇ *His ~ was tousled and he looked as if he'd just woken up.* | **cropped, long, short, shoulder-length** ◇ *She had shoulder-length black ~.* | **stray** ◇ *She pushed a stray ~ behind her ear.* | **thinning** | **body, facial, pubic** | **cat, dog, etc.** ◇ *The rug was covered with cat hairs.*

... OF HAIR **lock, wisp**

VERB + HAIR **have** ◇ *She had beautiful auburn ~.* | **lose** ◇ *He had turned forty and was beginning to lose his ~.* | **wear** ◇ *She wore her long ~ loose on her shoulders.* | **arrange, do, fix, tidy** (*esp. BrE*) ◇ *I don't like the way she's arranged her ~, do you?* ◇ *I'll be down in a minute, I'm just doing my ~.* ◇ *She showered, fixed her ~, and applied make up.* | **braid** (*esp. AmE*)**, plait** (*esp. BrE*)**, put up, tie back** ◇ *Why don't you put your ~ up for this evening?* | **brush, comb | shampoo, wash | cut, trim** ◇ *He went to the barber's to have his ~ cut.* […]

HAIR + VERB **grow** ◇ *Why don't you let your ~ grow?* | **curl** ◇ *His ~ curls naturally.* | **fall, flow, hang, lie, tumble** ◇ *Her blond ~ fell over her eyes.* | **gleam, glint, glisten, shine**

HAIR + NOUN **loss** ◇ *how to cope with ~ loss* | **salon, stylist** | **colour/color** | **accessory** ◇ *Her only ~ accessory was a headband.* | **extension** ◇ *a stylist specializing in ~ extensions* | **removal** ◇ *waxing, and other ~ removal methods available for men*

(OCD, McIntosh et al. 2009)

(7) comb *verb*

Comb is used with these nouns as the object: HAIR, INTERNET, WOOL, WRECKAGE

(OCD CD-ROM, McIntosh et al. 2009)

(8) excellent *adj.*

VERBS

appear, be, look, prove, seem, sound | consider sth

The school is considered ~.

ADV

most, really, truly

absolutely, quite

rather | generally

The meals are generally ~.

consistently, uniformly

The performances and recordings are uniformly ~.

apparently

potentially

otherwise

In an otherwise ~ issue, I found Creed's article very unconvincing.

PREPOSITION

at

Clancey was ~ at keeping the kids under control.

for

These potatoes are ~ for baking.

Excellent is used with these nouns: ACCOMMODATION, ACTING, ADVICE, AMENITY, ARTICLE, BALANCE, BARGAIN, BASE, BOOK, BUFFET, CAMOUFLAGE, CANDIDATE, CAST, CHAMPAGNE, CHANCE, CHARACTER, CHEF, CHOICE, COFFEE, COLLECTION, COMMAND, CONDITION, CONDUCTOR, COOK, COORDINATION, CREDENTIALS, CROP, CUISINE, DANCER, DESCRIPTION, DESIGN, DINNER, DISPLAY, DOCTOR, EDUCATION, EXAMPLE, EXCUSE, EYESIGHT, FACILITY, FIT, FOOD, FOUNDATION, FUN, GOAL, GOALKEEPER, GP, GRADE, GUIDE, HEALTH, HEARING, HOTEL, IDEA, ILLUSTRATION, IMPRESSION, INGREDIENT, INSTINCT, INTRODUCTION, INVESTMENT, JOB, LAWYER, LIBRARY, MATCH, MEAL, MEMORY, MODEL, MOTHER, MUSEUM, MUSICIAN, NEWS, OPPORTUNITY ,PACKAGE, PERFORMANCE, PHOTOGRAPH, PIECE, PLAY, PLAYER, POINT, PRESENTATION, PROGRESS, PROSPECT, PUB, QUALITY, RANGE, RATE, RECORD, RECORDING, RECOVERY, REFERENCE, REPORT, REPRODUCTION, REPUTATION, RESOURCE, RESTAURANT, RESULT, RETURN, REVIEW, RUN, SAVE, SCORE, SELECTION, SERVICE, SHAPE, SHOP, SHOT, SOURCE, SPEECH, START, STARTING POINT, SUBJECT, SUGGESTION, SUMMARY, TASTE, TEACHER, TRAINING, TRY, TUTORIAL, VALUE, VEHICLE, VISIBILITY, VISION, WEATHER, WIN, WINE, WORK

(OCD CD-ROM, McIntosh et al. 2009)

(9) impartial ADJ

not connected to or influenced by a particular person or group

● **ADV+ADJ** completely **absolutely, completely, entirely, strictly, totally** *Our aim is to provide completely impartial advice.*

▶ really **genuinely, truly** *We offer genuinely impartial advice.*

▶ in a way that relates to politics **politically** *A politically impartial civil service is a great national asset.*

▶ as some people believe **so-called, supposedly** *I cannot fully share his admiration for the supposedly impartial civil service.*

● **ADJ+N** person or group that judges **adjudicator, assessor, judge, panel** *Entries will be judged by a panel of impartial adjudicators.*

▶ person or group that settles disagreements **arbitrator, mediator, tribunal** *The mediator is impartial: he or she does not take sides.*

▶ person who makes sure people obey rules in sports **referee, umpire** *The lack of impartial referees allowed players to break the rules.*

▶ advice **advice, guidance** *The service offers impartial advice to new businesses.*

▶ journalism **journalism, reporting** *We are committed to honest and impartial reporting of the news.*

● and/or **balanced, fair ,objective, unbiased** *Members of the panel must be impartial and unbiased.*

(MCD, Rundell 2010)

(10) excellent *adj*

extremely good or of very high quality

NOUNS

excellent condition *The car is in excellent condition.*
excellent value *The hotel was excellent value.*
an excellent example *The palace is an excellent example of late 17th-century architecture.*
an excellent idea/suggestion/choice *I think the award is an excellent idea.*
an excellent job/piece of work *She does an excellent job of describing the problems that young people face.*
an excellent student/player/cook *Maria was an excellent student and passed all her exams easily.*
an excellent book/film/song *He wrote an excellent book about child psychology.*
excellent English/French/German etc. *The hotel staff all speak excellent English.*
an excellent article/report/paper *The paper has an excellent article on the current political situation in Greece.*
excellent food/meal *The food was excellent and I left a large tip.*
an excellent opportunity/chance *The meetings provide an excellent opportunity for discussion.*
an excellent result *Studies reported excellent results with the drug.*

ADVERBS

really excellent *His wife was a really excellent cook.*
absolutely excellent *I loved the speech – it was absolutely excellent.*
truly excellent *We increased our profit by 40% – a truly excellent performance!*
Don’t say ‘very excellent’.

(LCD, Mayor 2013)

In contrast, the MCD, the LCD and *Práctico* – all advertised as production dictionaries – do not have such a strict policy about including collocations in the lexical entry of the base or the collocate. This means that both collocates and lexical elements functioning as a base when combined with the headword are included in the same entry. For instance, the entries for the adjective *impartial* and *excellent* in the MCD (9) and the

LCD (10), respectively, contain both a list of collocate adverbs and a list of nouns constituting bases. One finds a similar situation in *Práctico* where, for instance, the entry of the verb *adquirir* 'acquire' lists both nouns (*fama* 'fame', *responsabilidad* 'responsibility' *derecho* 'right', etc.) and collocate adverbs or idioms (*a plazos* 'in instalments', *a granel* 'in bulk', etc.), see (11).

(11) adquirir v.

• CON SUSTS. **fama • notoriedad • popularidad • prestigio • credibilidad • reputación • respetabilidad • reconocimiento** *adquirir fama y reconocimiento mundial* **• respeto • renombre • consideración || deuda • compromiso • responsabilidad • obligación || costumbre • hábito • vicio || conocimiento • experiencia conciencia • cultura • formación • preparación • información • sabiduría • idea • vocabulario** *...para adquirir más vocabulario y mejorar la pronunciación* **bagaje • educación || capacidad • destreza • habilidad • competencia • soltura • versatilidad • seguridad • confianza • eficacia • elocuencia • talento • fluidez • práctica** *Para adquirir práctica comenzó a trabajar en...* **oficio || derecho • ciudadanía • nacionalidad • libertad • autonomía • independencia •** *adquirir independencia y autonomía en el trabajo* **legitimidad • inmunidad || importancia • relevancia • interés • peso • relieve • trascendencia • valor • gravedad • auge • grandeza • protagonismo || control • dominio** *adquirir dominio en un idioma* **autoridad || fuerza • poder • impulso • vigor • energía || sentido • significado • identidad • personalidad • entidad • voz || forma • espesor • masa • consistencia** *El argumento de la fiscal fue adquiriendo consistencia a medida que lo exponía* **resistencia• volumen • corporeidad • cuerpo • intensidad • dureza • firmeza || sida • gripe •** ***otras enfermedades*** **|| vivienda • coche • traje •** ***otros bienes materiales***

• CON ADVS. **a crédito** *los requisitos necesarios para adquirir a crédito una vivienda* **al contado • a plazos • a tocateja || a granel • a por mayor || a partes iguales • en exclusiva** *Las entradas para el espectáculo se adquieren en exclusiva a través de internet* **|| de golpe • progresivamente • sorpresivamente • gradualmente**

(*Práctico*, Bosque 2006)

Redes represents an approach which is different from other combinatorial dictionaries, since it provides more detailed descriptions of the combinatory properties of collocates – see (13) for the entry of the verb *apaciguar* 'ease' – and shorter concise lexical entries for bases of collocations – see (12) for the entry of the noun *enfado* 'anger'. This means that the dictionary emphasizes the decoding orientation. As it was discussed in the previous chapter (see 2.2.4.7and 2.3.2.1), Bosque justifies the structure of the dictionary claiming that collocational knowledge of speakers can be described as structured lists representing paradigms of bases which can be combined with a given collocate. Consequently, from his perspective, the layout of *Redes* is optimal for supporting collocation learning.

(12) **enfado** ♦ comprensible, descomunal, intenso, largo, malhumorado, mayúsculo^{14}, monumental^{77}, ostensible^{62}, pasajero^{38}, profundo, supino^{21}, tremendo, virulento, visible ♦ reacción (de) ♦ apaciguar, aplaca(se)^{5}, atemperar^{29}, causar, desencadenar(se)^{32}, dirigir (contra alguien), enterar(le) (a alguien), expresar, exteriorizar, hacer notar, írse(le) (a alguien), manifestar, mostrar, ocasionar, pasárse(le), (a alguien), provocar, remitir, sentir, sufrir, tener
□ Véase también: **cabreo, enojo, indignación**

(*Redes*, Bosque 2004b)

(13) **apaciguar** *v.* ▌Se construye frecuentemente con sustantivos que designan personas, animales o fuerzas naturales. También se combina con…

A SUSTANTIVOS QUE DESIGNAN SITUACIONES CONFLICTIVAS, SEAN DE NATURALEZA VIOLENTA, CONTROVERTIDA O POLÉMICA: 1 conflicto ++: ...puede jugar un papel decisivo para *apaciguar* el conflicto. LVE280395 **2 polémica ++:** Quizá esto contribuya a *apaciguar* la polémica de si los italianos descubrieron América. EPE020487 […]

B SUSTANTIVOS QUE DESIGNAN ELEMENTOS O FENÓMENOS NATURALES QUE SE ASOCIAN CON SITUACIONES DE AGITACIÓN O DE INESTABILIDAD. SE USAN MUY FRECUENTEMENTE EN SENTIDO FIGURADO: 11 agua +: …ha preferido no responderle para *apaciguar* las aguas ante el inminente inicio de la Eurocopa… EME080696 **12 tormenta:** Las declaraciones (…) *apaciguaron* la tormenta monetaria… LVE090395 […]

[…]

C EL SUSTANTIVO *ÁNIMO.* TAMBIÉN CON SUSTANTIVOS QUE DENOTAN DISGUSTO O IRRITACIÓN, A MENUDO MANIFESTADOS DE FORMA EXALTADA O TUMULTUOSA: 14 ánimo ++: Rafael Rey trata de *apaciguar* los ánimos en gesto entre ecuménico u desesperanzado. CAP190995 **15 ira +:** El policía fue detenido y será juzgado cuanto antes para *apaciguar* la ira de los trabajadores de la empresa de transportes. EME140296 […] **20 furia +:** Clinton, como una medida para *apaciguar* la furia de la UE, México y Canadá, pospuso por seis meses la aplicación del título tercero de dicha ley… EXC181296

[…]

F SUSTANTIVOS QUE DESIGNAN SENSACIONES O SENTIMIENTOS, MÁS FRECUENTEMENTE LOS QUE EXPRESEAN EL DESEO VIVO DE SATISFACER ALGUNA NECESIDAD: 34 sed ++: Ahí están los refrescantes datos para *apaciguar* la sed de los amantes de las estadísticas… EME020196 **35 pasión +:** Prefiero cometer estos errores que *apaciguar* mis pasiones LVE220796 […]

(*Redes*, Bosque 2004b)

Finally, DiCE, a dictionary whose theoretical foundation is constituted by ECL framework, focuses clearly on the direction of production when describing collocations. As follows from the definition of LFs, used to encode syntactic and semantic characteristics of collocations (see 2.3.3), the base of the collocation constituting a keyword to a LF figures as a headword in the dictionary, and the possible values of LFs are listed in the entry of the base. See (14) for the combinatorial information presented in the entry of the noun *enfado* 'anger'. With respect to information access in DiCE, it should be noted that regardless of the structure of the lexical entry, the online interface allows to query collocations containing a given base – an option not available in other electronic collocation dictionaries (see also 3.4.2.1.C).

When it comes to the organization of combinatory information within the lexical entry, as it was mentioned in the previous chapter (see 2.3), most dictionaries make use of two criteria to classify collocations: syntactic pattern and semantic content – an exception being the above mentioned short entries in *Redes*, which do not offer any classification. The primary organizing principle applied in all combinatorial dictionaries reviewed here corresponds to the syntactic pattern of the collocation. As we have seen in 2.3.1, dictionaries differ to some extent in the types of combinations they include, which affects the content of dictionaries in that some may offer combinations with a given syntactic pattern others choose to omit.

(14) **enfado 1** m. (Sentimiento)
enfado de individuo X con individuo Y por hecho Z

enfado + adjetivo

[–] **intenso** Magn

enérgico
ésta mostró su más enérgico enfado pues ni siquiera conocía a su futuro marido (web)

grande
Estos dos actos criminales recientes, preparados contra miembros de la Policía, despiertan gran enfado, preocupación y dolor en todos los dominicanos

mayúsculo
se enfrentaban ayer al tercer día consecutivo sin luz y su enfado crecía y se hacía mayúsculo (…)
Claro que esto no bastó para disipar su mayúsculo enfado (web)

[–] **más intenso de lo conveniente** Magn + Anti Ver

excesivo
Es posible, incluso, que si se le muestra un enfado excesivo o se le riñe, ni siquiera sea capaz de relacionar lo que ha ocurrido con su conducta (web).

[–] **poco intenso** Anti Magn

ligero
En mí la perplejidad dio paso a un ligero enfado

[–] **que dura mucho** Magn_temp

continuo
Comprobamos a diario la ausencia de comunicación en nuestras junglas urbanas, el habitual mal humor y el enfado casi continuo de todo el mundo

[…]

verbo + enfado

[–] **sentir ~** Oper_1

experimentar [(ART) ~]
Usar sólo mayúsculas equivale a gritar, experimentar enfado o descontento (web).

sentir [~]
Hacia sí mismo no sintió pena sino enfado

tener [ART ~]
Practicando la paciencia, el control del enfado llega a ser más fácil y, al tener menos enfado, uno quiere más a los demás en su vida diaria (web)

[–] **empezar a sentir ~** Incep Oper_1

coger [ART ~]
Es que así… con ese enfado que has cogido… (…)

[–] **continuar sintiendo ~** Cont Oper_1

conservar [ART ~]
Siempre me fue difícil conservar un enfado durante más de cinco minutos contra nadie (web)

[…]

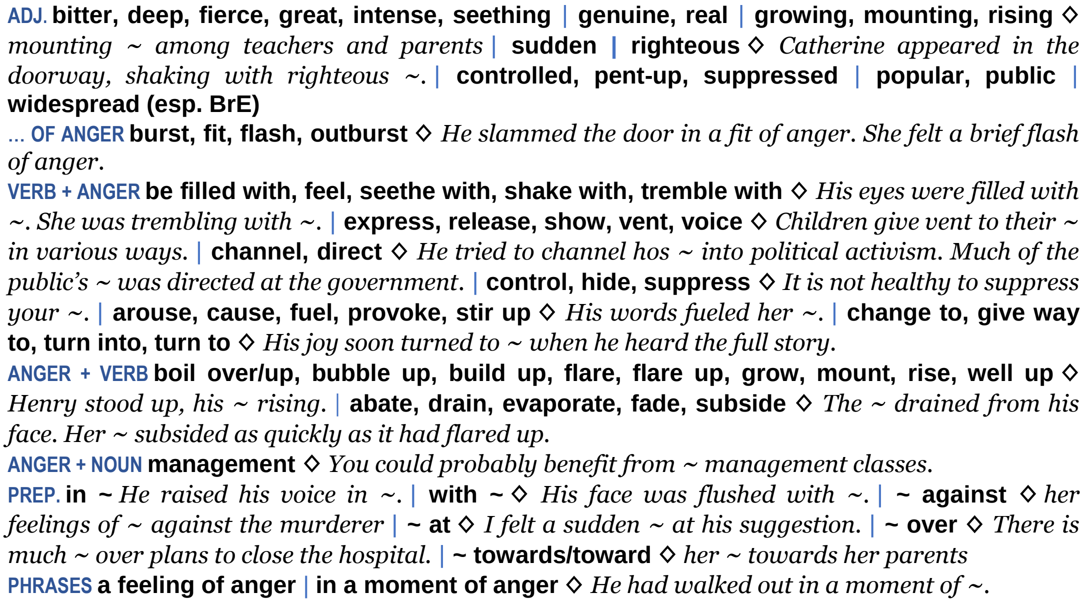
(adapted from DiCE, Alonso Ramos 2004)

Differences in how types of collocations are specified within the lexical entry may also affect the quality of the information from a usability point of view. For instance, as it can be observed in the sample dictionary entries shown in (15) and (16), both the OCD and *Práctico* list adjectives, nouns, verbs and prepositions in the entries corresponding to the nouns *anger* and *enfado* 'anger', respectively. However, while the OCD explicitly distinguishes between collocations corresponding to VERB+NOUN$_{\text{COMP}}$ and NOUN$_{\text{SUBJ}}$+VERB patterns, *Práctico* groups all verbal collocates under the heading CON VERBOS 'with verbs'. This is so, since this last dictionary seems to group collocations according to the part of speech of the components, rather syntactic pattern proper. I agree with Alonso Ramos (2008, 1218) and Buendía Castro and Faber (2014, 221–222), that the lack of distinction in the latter dictionary between cases when the noun appears as subject and object of the verb in a combination can be confusing for the language learner.

Within each main combinatory group corresponding to a given syntactic pattern, most dictionaries establish subcategories according to the meaning of the items listed (see 2.3.2.1). The semantic criteria used to identify each subgroup are, however, often not explicitly indicated. Such is the case in the OCD or in *Práctico*, where collocates are organized in groups according to "semantic proximity". In (15), for instance, one can deduce that the first group of verbs combining with *anger* (*be filled with, feel, shake with, tremble with*) make reference to the FEELING OR CORPORAL MANIFESTATION RESULTING FROM FEELING ANGER, while those in the second group (*express, release, show, vent, voice*) refer to the EXPRESSION OF ANGER and those in the fourth group (*control, hide, suppress*) all have meanings related to NOT EXPRESSING ANGER. In (16), the first three groups of adjectives combining with *enfado* 'anger' can be characterized by the general meanings of INTENSE (*tremendo* 'huge', *descomunal* 'huge' *monumental* 'monumental'), MANIFEST (*visible* 'visible', *ostensible* 'obvious') and LOW INTENSITY (*pequeño* 'small', *ligero* 'light'). While the OCD and *Práctico* do not include semantic labels at all, the LCD indicates the meaning of less transparent combinations, such that e.g. in the case of the noun *anger* the meanings of the collocation groups *feel anger* and *express/show/vent anger* are not specified, while, the meaning of groups *cause/provoke/arouse/stir up anger* and *fuel anger* are explicitly indicated as 'make people angry' and 'make people even more angry', respectively, see (17). In comparison, in both the *MCD* (9) and DiCE (14), approximate meanings of combinations included in each subgroup are systematically indicated, which certainly facilitates the interpretation of the content of the dictionary to its users.

Finally, *Redes* (Bosque 2004b) also contains explicit semantic labels, since, as we have seen, in a way, semantic classification constitutes the aim of the dictionary (see 2.2.4.7 and 2.3.2). Consequently, in the detailed entries of collocates, bases are grouped into what Bosque calls *lexical classes*, which are assumed to contribute to the description of the meaning of the collocate. As we can observe in (13), the Spanish verb *apaciguar* ‘ease’ typically co-occurs with, for instance, a) NOUNS DESIGNATING CONFLICTIVE SITUATIONS, BEING VIOLENT, CONTROVERSIAL OR POLEMIC, b) NOUNS DESIGNATING NATURAL ELEMENTS OR PHENOMENA ASSOCIATED WITH UNREST OR INSTABILITY, OFTEN IN A FIGURATIVE SENSE or c) NOUNS DENOTING DISGUST, IRRITATION, OFTEN MANIFESTED LOUDLY AND WITH EXAGGERATION, etc.

(15) anger *noun*



ADJ. **bitter, deep, fierce, great, intense, seething** | **genuine, real** | **growing, mounting, rising** ◇ *mounting ~ among teachers and parents* | **sudden** | **righteous** ◇ *Catherine appeared in the doorway, shaking with righteous ~.* | **controlled, pent-up, suppressed** | **popular, public** | **widespread (esp. BrE)**

... OF ANGER **burst, fit, flash, outburst** ◇ *He slammed the door in a fit of anger. She felt a brief flash of anger.*

VERB + ANGER **be filled with, feel, seethe with, shake with, tremble with** ◇ *His eyes were filled with ~. She was trembling with ~.* | **express, release, show, vent, voice** ◇ *Children give vent to their ~ in various ways.* | **channel, direct** ◇ *He tried to channel hos ~ into political activism. Much of the public's ~ was directed at the government.* | **control, hide, suppress** ◇ *It is not healthy to suppress your ~.* | **arouse, cause, fuel, provoke, stir up** ◇ *His words fueled her ~.* | **change to, give way to, turn into, turn to** ◇ *His joy soon turned to ~ when he heard the full story.*

ANGER + VERB **boil over/up, bubble up, build up, flare, flare up, grow, mount, rise, well up** ◇ *Henry stood up, his ~ rising.* | **abate, drain, evaporate, fade, subside** ◇ *The ~ drained from his face. Her ~ subsided as quickly as it had flared up.*

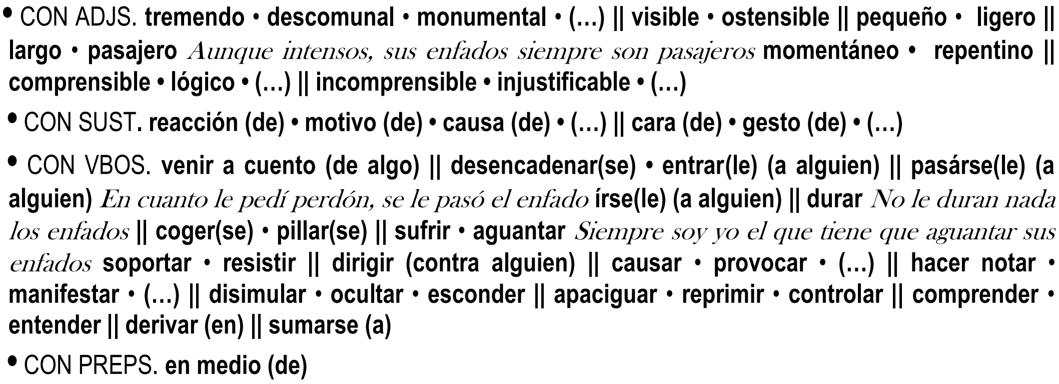
ANGER + NOUN **management** ◇ *You could probably benefit from ~ management classes.*

PREP. **in ~** *He raised his voice in ~.* | **with ~** ◇ *His face was flushed with ~.* | **~ against** ◇ *her feelings of ~ against the murderer* | **~ at** ◇ *I felt a sudden ~ at his suggestion.* | **~ over** ◇ *There is much ~ over plans to close the hospital.* | **~ towards/toward** ◇ *her ~ towards her parents*

PHRASES **a feeling of anger** | **in a moment of anger** ◇ *He had walked out in a moment of ~.*

(OCD, McIntosh et al. 2009)

(16) enfado s.m.



- CON ADJS. **tremendo • descomunal • monumental • (...) || visible • ostensible || pequeño • ligero || largo • pasajero** *Aunque intensos, sus enfados siempre son pasajeros* **momentáneo • repentino || comprensible • lógico • (...) || incomprensible • injustificable • (...)**
- CON SUST. **reacción (de) • motivo (de) • causa (de) • (...) || cara (de) • gesto (de) • (...)**
- CON VBOS. **venir a cuento (de algo) || desencadenar(se) • entrar(le) (a alguien) || pasárse(le) (a alguien)** *En cuanto le pedí perdón, se le pasó el enfado* **írse(le) (a alguien) || durar** *No le duran nada los enfados* **|| coger(se) • pillar(se) || sufrir • aguantar** *Siempre soy yo el que tiene que aguantar sus enfados* **soportar • resistir || dirigir (contra alguien) || causar • provocar • (...) || hacer notar • manifestar • (...) || disimular • ocultar • esconder || apaciguar • reprimir • controlar || comprender • entender || derivar (en) || sumarse (a)**
- CON PREPS. **en medio (de)**

(*Práctico*, Bosque 2006)

(17) anger *n*

a strong feeling you have when someone has done something bad

[…]

VERBS + anger

feel anger *He felt no anger, just sorrow.*

express/show your anger *also* **vent your anger** *formal Demonstrators expressed their anger by burning American flags.*

cause/provoke/arouse/stir up anger (=make people angry) *The referee's decision provoked anger among the fans.*

fuel anger (=make people even more angry) *The announcement fueled public anger against the government.*

explode with anger (=suddenly express great anger) *When he found he had been cheated, he exploded with anger.*

control/contain your anger *I could not control my anger any longer.*

hide your anger *For a second she was unable to hide her anger.*

anger + VERBS

sb's anger goes away/subsides/fades (=it stops) *I counted to ten and waited for my anger to go away. | His anger slowly subsided.*

sb's anger grows/rises *Her anger and resentment grew as she drove home. | Thinking this, he felt his anger rise again.*

anger boils up/wells up inside sb (=someone suddenly becomes very angry) *She could feel the anger boiling up inside her.*

ADJECTIVES

deep/great/fierce anger *There is deep anger against the occupying forces.*

growing/rising/mounting anger *There is growing anger among drivers over the rise in fuel prices.*

widespread anger (=among many people) *The decision to build the airport has provoked widespread anger.*

real anger *There is real anger about the amount of money that has been wasted.*

public/popular anger *By now public anger in the US was mounting.*

suppressed/pent-up anger (=that you have tried not to show) *Her voice shook with suppressed anger.*

righteous anger (=anger felt when you think something should not be allowed to happen) *The speech was full of righteous anger against the West.*

[…]

(LCD, Mayor 2013)

C. Presentation of collocations in collocation dictionaries

The previous section considered the presentation of combinatorial information in collocation dictionaries from two main perspectives: whether they are oriented to decoding or encoding, and how combinatorial information is organized within the lexical entry. We have seen that the second implies describing two important characteristics of collocations, their syntactic pattern and their meaning. However, as it will be discussed in this section, there is certain additional information that is necessary in order for a language learner to be able to use a combination appropriately when producing a text.

Heid (2004, 730–731) proposes a “maximalist” approach to the description of collocations, following which lexical combinations should be described in as much detail as single lexical items. The series of phenomena that should be attended to by such description, according to the author, includes 1) lexical combinatorics, 2) morphosyntax,

3) syntactic subcategorization, 4) semantic properties and 5) pragmatic properties (see Table 8). Note that the first type of information, referring to lexical combinatorics, which involves the specification of the number of components and the indication of whether a given set of collocates corresponds to an open or closed list is not relevant in the case of the notion of collocation adopted in the present thesis, since, according to this, collocations are always composed of two lexical units, and collocate lists are always closed – since they are considered to be arbitrary (see 2.2.5). The remaining information types, nevertheless, refer to phenomena that a language learner has to be familiar with in order to use a given combination in a native-like manner, and, as it is demonstrated by the results of the corpus study presented in the following chapter, are indeed often problematic for non-native speakers. Consequently, it is important to explore whether and to what extent these types of information are represented in collocation dictionaries. For this, I will again rely on the five collocation dictionaries mentioned in the previous section: the English language dictionaries OCD, MCD, LCD and the Spanish combinatory dictionaries *Redes*, *Práctico* and DiCE.

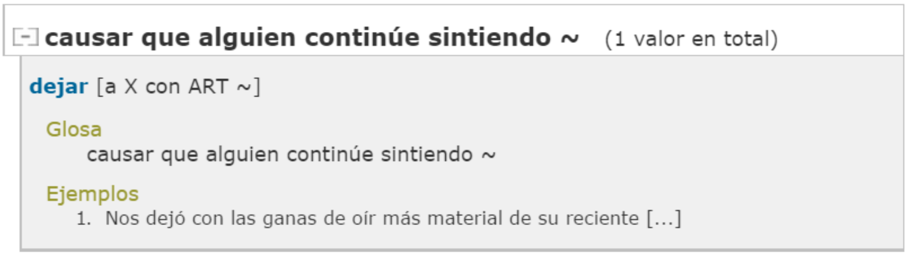
Level	**Phenomenon**	**Examples**
Lexical Combinatorics	Number of lexemes Open/closed collocate list	*poner+atención* 'pay attention' *enfado {tremendo, descomunal, monumental…}* '{tremendous, colossal, monumental} anger'
Morphosyntax	Noun: singular/plural Modifiability of the noun Determination of the noun	*tener tendencias ADJ* 'have ADJ tendencies' *restaurar la confianza a X* 'restore sb's confidence to X'
Syntactic subcategorization	Verb valency Noun valency	*dar una sorpresa a alguien* ' surprise somebody' *albergar la esperanza de* 'cherish the hope of'
Semantic properties	Synonymy	*profesar/sentir admiración* 'profess/feel admiration'
Pragmatic properties	Diasystematic marks Frequency in a corpus	Style, geographic use

Table 8: Summary of types of information required to describe a collocation, adapted from Heid (2004: 731)

Morphosyntactic information mentioned by Heid (2004) is often not indicated explicitly in dictionaries. For instance, in the case of the combination *make the bed* or its Spanish equivalent *hacer la cama* involving determination, there is no overt indication of the obligatory use of the article in the OCD or the *Práctico* respectively. The LCD provides more complete information since it spells out each collocation lemma in its full from, instead of providing a mere list of collocates, as other dictionaries do. Furthermore, in the case of Spanish, it is notable that while *Práctico* indicates the gender of the noun constituting the headword, it does not include such information in the case of collocate nouns, such as with *reacción (de)* ‘reaction of’, *motivo (de)* ‘cause of’, *causa (de)* ‘cause of’ co-occurring with *enfado* (see (16)).

In relation to syntactic information, it is the valency or government pattern of the verb that is most often indicated explicitly, typically in the lemmatized form of verbal collocates. In (15), we can observe, for instance, that the OCD indicates governed prepositions in the case of VERB+NOUN$_{\text{COMP}}$ combinations with the noun *anger* (*be filled with, seeth with, shake with, tremble with*), similarly to e.g. the case of *Práctico*, where some of the verbal collocates of *enfado* ‘anger’ are indicated as *entrar(le) (a alguien)* lit. ‘enter (somebody)’ *dirigir (contra alguien)* lit. ‘direct (against somebody)’ (see (16)), in the same manner as in the short entry of the same noun provided in *Redes* (see (12)), and the LCD, which includes lemmatized forms of collocations such as *show your anger* and *anger boils up inside sb* (see (17)). The valency of nouns constituting the headword is typically given as a type of combination in its own right, as e.g. in the OCD where *anger against, anger at, anger over* and *anger towards/toward* are listed as prepositional combinations.

From among the collocation dictionaries mentioned so far, DiCE is in general the most consistent in including overt information concerning the characteristics of collocations. This is so since it is conceived within the framework of the ECL placing considerable emphasis on the explicit description of government pattern, as well as other types of information relevant in usage in the dictionary. As it can be seen in Figure 1, in the case of the combination of the noun *gana* and the verb *dejar*, the dictionary specifies that the verb requires a direct object, *a X* standing for *a alguien* ‘somebody’ and takes the base noun as a prepositional complement, introduced by the preposition *con* ‘with’. It is also indicated that the base noun is preceded by a determinant (*ART*). At the same time, the government pattern of the headword *gana* constituting the base is described in detail in a section the user can access from the main entry of the lexical unit itself (see Figure 2); for more detail on the description of syntactic characteristics of lexical units in DiCE see Alonso Ramos (2003, 18–19).



causar que alguien continúe sintiendo ~ (1 valor en total)

dejar [a X con ART ~]

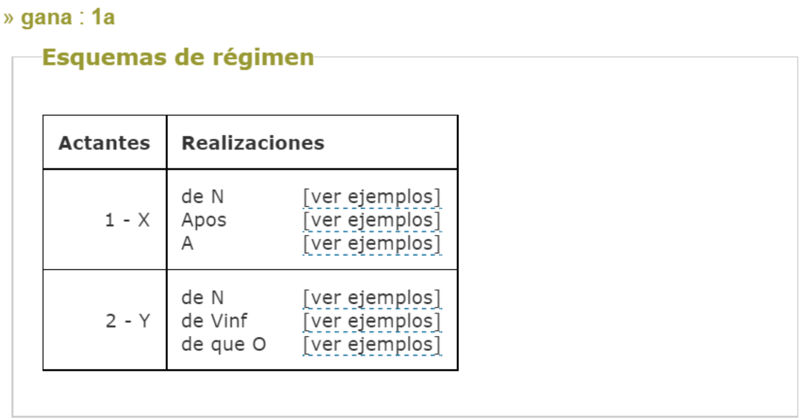
Glosa

causar que alguien continúe sintiendo ~

Ejemplos

1. Nos dejó con las ganas de oír más material de su reciente [...]

Figure 1 Lexical entry for the collocation *dejar con las ganas* lit. ‘leave with the wishes’ ‘make somebody want to do something’ in DiCE



» gana : 1a

Esquemas de régimen

Actantes	Realizaciones	
1 - X	de N Apos A	[ver ejemplos] [ver ejemplos] [ver ejemplos]
2 - Y	de N de Vinf de que O	[ver ejemplos] [ver ejemplos] [ver ejemplos]

Figure 2 Government pattern of the headword corresponding to the lexical unit *gana 1a* in DiCE

Patterns of use involving both syntactic and morphosyntactic information can also be observed in usage examples, when present, although collocation dictionaries differ to a great extent regarding how systematically they provide these. While the OCD and *Práctico* are the least consistent in offering usage examples, the MCD and the LCD provide at least one example in the case of each semantic set of combinations. DiCE as well as *Redes*, in the case of its long entries, are the most consistent in that they provide examples in the case of every single collocation, however, these are taken directly from language corpora, and are not adapted to language learners’ needs.

Information provided regarding the meaning of combinations was discussed in detail in the preceding section, where it was noted that in some dictionaries such as the OCD and *Práctico* the semantic content of collocations is not explicitly indicated, in contrast to the LCD, which does provide an overt description of meanings in certain cases, and the MCD and DiCE, which include semantic glosses in all cases.

With regard to the pragmatic properties mentioned by Heid (2004), frequency information – considered to be relevant from a pedagogical point of view – is offered in *Redes* and in DiCE. In the case of the latter, data on frequency of occurrence has been obtained for both the bases and the combinations, and is explicitly shown in the user interface in the case of bases, while collocations belonging to the same combinatory group can be ordered from higher to lower frequency (Vincze and Alonso Ramos 2013).

While none of the above mentioned collocation dictionaries contain usage labels, the potential of adding such information to the collocation entries of DiCE was explored by Vázquez Veiga (2014), who proposes that diaphasic and diaevaluative marking of the combinations should be attended to through labels such as *formal, informal, vulgar, euphemistic* and *pejorative*. Empirical evidence to the register-specific nature of collocations is provided by e.g. Corpas Pastor (2015), who compared the range verbs used in VERB+NOUN$_{\text{COMP}}$ combinations in specialized medical corpora and general reference corpora of English and Spanish.

D. Summary of the characteristics of collocation dictionaries

The previous two sections dealt with the presentation of combinatorial information in six specific collocation dictionaries in detail, in order to provide an overview of the notions involved in the lexicographical description of collocations. In order to round up this discussion, in Table 9 I provide a summary of the main characteristics of these dictionaries. Since so far I have focused on the manner collocations are represented, an important aspect referring to the medium of dictionaries itself, i.e. to whether they constitute paper editions or are published in electronic format, has been neglected.

From among the dictionaries presented here DiCE is the only one that was originally conceived as an electronic dictionary, while both the OCD and the LCD have electronic editions in CD-ROM and online format, respectively. Nevertheless, when it comes to providing different access paths to combinatorial information, it is only DiCE which attempts to exploit more fully the possibilities of the electronic medium. While, the electronic versions of the LCD and the OCD merely allow accessing entries in the same way as the print version of the dictionaries, i.e. through the headword, *DiCE* provides a number of different search options, which allow the user to query the dictionary database in a more dynamic way. Thus, as it was mentioned earlier, it is possible, for instance, to retrieve all instances of a given collocate through a specific search option. The different search options offered by DiCE, their utility and degree of ease of use will be further considered in the usability study described in Chapter 5 (see 5.2). An advantage of the CD-ROM version of the OCD is that it provides access to the monolingual dictionary entry of each word included through hyperlinks.

	ORIENTATION OF ACCESS	CLASSIFICATION	OTHER INFORMATION	ADDITIONAL TYPES OF ACCESS
OCD	Base CD-ROM: Base and collocate	1. Syntactic pattern 2. Semantic – not explicit	Verb valency in collocate lemmas; Governed prepositions as an individual collocation type Examples of use – not systematic	CD-ROM Only access through headword Hyperlinks to definitions
MCD	Base and collocate – in one lexical entry	1. Syntactic pattern 2. Semantic – explicit	Valency in collocation patterns (e.g. V+*with*+N) Examples of use in each semantic set	N/A
LCD	Base and collocate – in one lexical entry	1. Syntactic pattern 2. Semantic – not explicit	Whole collocation lemmas (e.g. *make the bed*) Examples of use in each semantic set Thesaurus	Online access Only access through headword
Redes	Collocate in long entries Base in short entries	Long entries: 1. Part of speech 2. Semantic – explicit Short entries: no classification	Long entries: Documented corpus examples with each combination Frequency of use Short entries: Verb valency in collocate lemmas	N/A
Práctico	Base and collocate – in one lexical entry	1. Part of speech 2. Semantic – not explicit	Verb valency in collocate lemmas Examples of use – not systematic	N/A
DiCE	Base	1. Syntactic pattern 2. Semantic - explicit	Information on government pattern and use of articles Examples of use from corpus Lexical frequency of bases	Free online access Multiple advanced search options

Table 9 Summary of main characteristics of collocation dictionaries

E. Using dictionaries for learning collocations

The six dictionaries described in the previous sections are all indicated to be used in foreign language teaching. In fact, a number of authors propose specific learning activities and/or exercises using dictionaries. In the EFL context, Hill et al. (2000, 99–115) proposes a good number of learning activities involving the use of collocation dictionaries, while McGee (2012, 354–359) designs inductive learning activities based on a number of specific English collocation dictionaries, including the OCD and the MCD. As for the field of SFL, Hoyos Puente and Villar Díaz (2006), Ruíz Martínez (2006) and Serrano Dolader (2007) all highlight the pedagogical value of the lexical description provided in *Redes*.

Higueras García (2006, 52–54) offers a series of learning activities using this dictionary; in a subsequent work the same author proposes two activity sequences using *Práctico* (Higueras García 2008), finally Ferrando Aramo (2009) develops an online teaching sequence which prompts language learners to work with DiCE. In addition, noting that none of the existing Spanish combinatorial dictionaries was designed specifically for language learners, Ferrando Aramo (2012) proposes a bilingualized collocation dictionary format especially suited for Italian learners of Spanish, which allows filtering combinatory information according to proficiency levels and conceptual fields (adopted from PCIC – see above).

The above mentioned pedagogical proposals are all founded on the assumption that the use of dictionaries can contribute to enhancing language learners' collocation competence. When assessing this idea, Laufer (2010, 32) formulates three conditions which need to be fulfilled in order for actual collocation learning to take place. Firstly, dictionaries have to provide the target collocations and present them at the adequate headwords where users can find them. Secondly, learners have to decide to look for the collocations, for which they have to be aware of their existence and the difficulties they pose in a foreign language. Thirdly, leaners have to remember the collocations they looked up after having completed the task. Accordingly, Laufer and other authors have carried out empirical studies in order to verify whether and how learners look up collocations, if they are able to find them and if learning takes place. These studies mainly concentrate on the use of MLDs – it is argued, e.g. by Laufer (2010, 32) that learners most often have access to and use this type of dictionaries, while Handl (2009, 71) claims that they prefer to use one all-purpose reference tool –, while only a handful of papers investigate the use of collocation dictionaries.

Dziemianko (2010) and (2011) report on studies whose aim was to compare the usefulness of the paper and electronic versions of two specific MLDs, the *Collins COBUILD Advanced Dictionary* (COBUILD, Sinclair 2008) and the *Longman Dictionary of Contemporary English 5th edition* (LDOCE5, Mayor 2009), respectively. The studies evaluated participants' performance with the dictionaries in a receptive and a productive task, and also assessed the role of dictionary form in the retention of target items, which was measured in an unannounced post-test. The production task devised by the author consisted of the completion of multiword expressions – mostly combinations of prepositions and nouns (e.g. *on the blink, in cahoots with, up the creek*) – whose presentation in the dictionaries is similar to that of collocations. In the case of the first experiment, involving COBUILD, participants using the electronic version of the dictionary were found to perform significantly better both in the case of the reception and production tests, as well as the retention test. From this the author concluded that the electronic dictionary was a better learning tool probably because "the visual impact created by the electronic dictionary and the prominent position of a headword on the computer screen can attract more attention than a printed page", while "the ease of look-up and the saliency of an entry on the computer screen are more beneficial to the learning process than the effort put into the extraction of relevant information from a paper dictionary" (Dziemianko 2010, 265–266). The attempt to replicate the study with LDOCE5, however, did not result in similar findings, as the medium of the dictionary was not found to have a statistically significant effect on participants' performance. Dziemianko (2011, 97) attributed this to the fact that the colorful widgets found on the LDOCE5 online interface may have made dictionary content less salient to users.

In a more recent study, Dziemianko (2014) addressed the question of how the presentation and placement of collocations in the lexical entry of MLDs affect their use and retention. The test administered to participants consisted of gapped sentences where the collocates had to be supplied with the help of systematically manipulated dictionary entries. It was found that participants performed significantly better on the production test when collocations were given in bold in the dictionary entries – either embedded in examples or before them, while the placement of collocations in a final position was found significantly more effective than the entry initial position. The same effects were found in the case of the retention test, although the positive effect of entry final placement was not statistically significant.

Laufer (2010) tested the usefulness of different MLDs and a bilingualized dictionary – which included the translation of the headword in the learners' L1 – in a production test as well as a retention test involving verb+noun collocations (e.g. *put pressure, take measures, get the message*). The author found that although participants performed significantly better when they could consult the dictionary entries than in the pre-test done without them, they in fact did not find all collocations that were included in the dictionaries. She also observed that participants did not look up many collocations because they were not aware of the fact that they were unfamiliar with them, while most collocations they managed to find were not retained a week later when the retention test was administered. Laufer interpreted her results as alarming, claiming that they point at the importance of raising learners' awareness to the existence of collocations, as well as training them in the use of dictionaries.

Two of the studies focusing on the use of collocation dictionaries dealt with the *Oxford Collocations Dictionary 1st edition* (OCD1, Crowther et al. 2002). The first of these, Komuro (2009) aimed at finding out how successfully Japanese EFL students could look up verb+noun, adjective+noun and preposition+noun collocations to complete gapped sentences constituting the translation of Japanese sentences also provided in the test. She concluded that her participants could in general successfully interpret the part of speech groupings of collocations offered in the dictionary, although, on occasions they failed at determining the syntactic pattern of the target collocation. This, in my opinion, might have resulted from the low proficiency level, and consequently, weak sense of syntactic structures of the participants, after all, their task was to complete a fill-in-the-blank type exercise. Although, since the author suggests that combinations whose syntactic pattern was incongruent with their L1 translation seemed to be problematic, it may also be the case that participants relied too much on the structure of the original Japanese sentences when trying to convey their meaning in English. The author also noted that the semantic sets offered by the dictionary resulted confusing to the students, who were overwhelmed with the large amount of collocates presented together, and would have needed more example sentences or explicit meaning indications to interpret them.

The second study, Lew and Radłowska (2010) compared intermediate EFL learners' performance in finding collocations in the OCD1 and a MLD, the *Longman Dictionary of Contemporary English 4th edition* (LDOCE4, Bullon 2004). The test consisted of 13 gapped sentences in which members of different types of collocations (e.g. verb+noun, noun+noun, adverb+adjective, etc.) had to be supplied. Surprisingly, participants were found to perform better with the LDOCE4, although the difference between the two dictionaries was not statistically significant. The authors' observations regarding the OCD1 are similar to that of Komuro's (2009), in that they claim that most problems encountered by learners seemed to stem from lack of comprehension: the vague proximity of meaning groupings in the OCD1 resulted in the confusion of collocates, while the lack of examples or explicit indication of the meaning of combinations seemed to affect performance negatively.

In a study involving Spanish combinatory dictionaries, Alonso Ramos (2008) assessed the use of *Práctico* and DiCE in a multiple choice test where participants had to select the suitable verb+noun combination for a given context. The results of her experiment showed that both dictionaries proved useful in that they helped participants to improve their performance with respect to the pre-test administered without a dictionary, however, the use of the dictionaries did not always lead to positive results. In certain cases, when the correct answer was not to be found in the dictionaries, participants seemed to succumb to the authority of the reference tools in that they changed their correct answer provided in the pre-test to an incorrect one, consisting in a collocation they managed to find in one of the dictionaries. Other cases highlighted the importance of instructing learners in the use of dictionaries, since, for instance, in certain cases participants did not seem to be able to interpret certain information, such as the indication of the government pattern [~ *a/en* X] in the case of the combination *provocar celos en alguien* lit. 'provoke jealousy in somebody'.

The empirical studies mentioned above assessed the usefulness of dictionary products. Heid (2011) adopts the term *usability testing*, commonly used in information science, to describe studies with a similar aim, although limiting his scope to electronic dictionaries, which, according to him can be seen and tested as software tools. As Heid and Zimmerman (2012, 665) remark, the criteria typically applied in usability studies are "conformity to user expectations, consistency, error tolerance, learnability and memorability". We have to recognize that most of these were targeted in the studies that have been reviewed here, since they all aimed at assessing one or more dictionaries as to the degree of successful use in a given task, and/or evaluating certain aspects of the presentation of lexicographical information. This demonstrates that usability testing in effect is not solely applicable in the case of electronic, but also to paper dictionaries.

Heid (2011) also highlights the importance of usability studies as part of the development process of new lexicographical tools. This is in line with some of the studies presented here, given they dealt with lexicographical products that are not as yet finalized. Dziemianko (2014) created systematically manipulated lexical entries to measure the impact of certain variables related to the presentation of combinatorial information, while both Laufer (2010) and Alonso Ramos (2008) compared the use of a commercial dictionary with that of an ongoing dictionary project. In 5.2, I describe a study applying usability testing methodology in order to evaluate the success of user interactions with the current version of the online DiCE interface.

To sum up, the results of empirical research reviewed here provide important pointers as to improving accessibility of collocational information in dictionaries in general: more straightforward access, clearer and better structured lexical entries, and highlighting of combinations in the dictionary entry and in examples seem to help identify the target item and promote retention in the case of MLDs, while the inclusion of L1 translation equivalents and easily interpretable morphosyntactic information also contribute to the success of dictionary use.

These studies also give a hint regarding the adequacy of the way collocations are presented within the lexical entries of collocation dictionaries. Both Komuro (2009) and Lew and Radłowska (2010) found that the way collocates are grouped in semantic sets according to proximity of meaning is confusing to users, who often need an explicit indication of meaning. As for the classification of combinations according to syntactic pattern, Komuro (2009) concluded that, in general, it did not cause much difficulty to learners to locate desired combinations, however, combinations whose structure is incongruent with L1 equivalents may be problematic. This raises the question of whether the primacy of syntactic pattern or part-of-speech offers the most efficient access route to collocations in all cases, an issue which is further considered in the following subsection.

Finally, it is important to note that dictionary use studies in general provide very little insight into the effectiveness of the organization of collocation dictionaries. Firstly, only a considerably low number of empirical studies have focused on this type of dictionaries, and, secondly, the tasks used in these studies, consisting of the completion of gapped sentences with collocates, are tailored to the content of the dictionaries and, moreover, prompt the exact look-up mechanism or search strategy these are designed for. In this sense it would be interesting to test collocation dictionaries in a genuine production task, such as essay writing. The relationship between dictionary structure and search strategies is further considered in the following subsection.

F. Towards a more dynamic access in collocation dictionaries

The preceding sections discussed how combinatorial information is presented to the user in some of the most common combinatorial and collocation dictionaries of English and Spanish, and reviewed a number of studies which aimed at assessing the relationship between certain aspects of the presentation of combinatorial information and the degree of successful dictionary use on the part of the language learner. Clearly, the way word combinations are organized in dictionary entries has direct consequences on the types of look-ups dictionary users are able to carry out. Production-oriented dictionaries, which provide a list of collocates in the entry of the base of a collocation, are in general designed to be used in lookup situations which can be illustrated by the following question: 'What adjectives can I use to speak about an increase of the intensity of *anger*?' Note that the formulation of this question implies that the dictionary user first has to determine the part of speech of the item and/or the syntactic pattern of the expression they are searching for. Thus, with the help of, for instance, the OCD (see corresponding the lexical entry in (15)), they would be able to find the adjectives *growing, mounting* and *rising*. However, in order to identify other expressions, such as *fuel somebody's anger* or *anger grows/mounts/flares up,* which do not match the pre-supposed syntactic pattern, but express similar meanings, they would have to read through the whole length of the dictionary entry.

Combinatorial information in the dictionary can be organized in an alternative way which allows access to the data through different search options. If syntactic and semantic classifications of collocations are independent of each other, the user can decide whether they want to search for combinations according to one or the other criterion in each look-up. Through applying this model, besides the access route exemplified above, a collocation dictionary can also allow the user to delimit the meaning of the desired combination first, and then select from among a number of (nearly) synonymous expressions the one which fits a given context. This type of look-up – not supported by any of the above mentioned dictionaries – can be illustrated in the following way: 'What expressions can I use to speak about an increase of the intensity of *anger*?', and would render as a result a list containing all expressions mentioned in the previous paragraph.

Jousse's (2010) proposal represents exactly this innovative approach. Importantly, dissociating semantic and syntactic classification of combinatorial data, and allowing semantic searches requires the implementation of a systematic semantic typology throughout the dictionary. That is why this scholar developed a comprehensive typology of collocations (presented in 2.3.2.1) in order to enhance search possibilities in an electronic lexical database, catering to production-oriented user needs in the above mentioned way. Once this type of semantic classification is implemented, a novel access option would enable the user to search for collocates used to convey a given meaning when combined with a specific term. For instance, one could easily retrieve the collocates that can be used to express the idea ‘start’ when combined with *miedo* ‘fear’: *entrar(le)* [*miedo*] ‘fear enters (sb)’ and *coger* [*miedo*] ‘to catch fear’; with *amistad* ‘friendship’: *entablar* [*amistad*] ‘to start a friendship’, [DET *amistad*] *nace* ‘friendship is born’; or with *amor* ‘love’: [DET *amor*] *surge* ‘love emerges’, independently of their syntactic pattern.

This meaning oriented search path is further discussed in L’Homme and Leroyer (2009) as well as in Jousse et al. (2011). The latter describes the implementation of the proposal in order to enrich the presentation of combinatory information in *DiCoInfo* (L’Homme 2009), a terminological dictionary specialized in the field of internet and computation. Collocations are reorganized in groups and subgroups representing typical meanings expressed by lexical combinations relative to the semantic field dealt with by the dictionary. Consequently, verbal collocates such as e.g. *récupérer* ‘recover’, *restaurer* ‘restore’, *modifier* ‘modify’, *éditer* ‘edit’, noun collocates as e.g. *modification* ‘modification’, *édition* ‘edition’ and the adjective collocate *éditable* ‘editable’ of the noun *fichier* ‘file’ are all included in the semantic class *utiliser/faire fonctionner* ‘use/operate’.

While the authors show the tentative results of a usability study carried out with the dictionary interface, no clear conclusions are drawn as to the effective utility of the semantic classification applied. In any case, it is more likely that both classification of combinations on the lexicographers’ part, and interpretation of semantic classes by the users are less problematic in the case of a restricted domain dictionary, than in the case of a general combinatory dictionary. Nevertheless, since I find this approach especially appealing for the case of collocation learning, given it potentially allows a language learner to discover a fuller repertoire of collocations expressing similar meanings in the dictionary, I propose to adopt it in the case of the online collocation learning tool described in Chapter 6.

3.4.2.2 Language corpora as collocation learning resources

Pedagogical applications of corpora are commonly classified as indirect and direct (see e.g. Römer 2011, 206–207). Indirect applications refer to the use of corpora by researchers and teachers in syllabus design or the development of teaching materials. Direct applications involve the language learner and teacher in actively working with corpora and concordances. This latter type of applications is also commonly referred to by the term *data-driven learning* (DDL), generally associated with Tim Johns' pioneering work starting in the 1980's (see Johns 1986).

In what follows I briefly review some of the arguments in favor of using corpora in language teaching in general, and more specifically for teaching collocations. After this, I describe a number of studies that have attempted to examine students' use of corpora in language learning tasks.

A. Rationale for using corpora for learning collocations

Advocates of incorporating corpora in the foreign language classroom have put forward a number of beneficial aspects of this methodology (see e.g. Yoon and Hirvela 2004). A major argument states that corpora are considered to be representative of language as used naturally – in a non-classroom context – in terms of the frequency of occurrence of different linguistic phenomena. Working with corpora is also said to increase learners' opportunities of contact with texts in the foreign language, while it is claimed that exposure to authentic texts contributes to improving the understanding of how specific lexical items are used in particular contexts. Ultimately, the use of corpora is claimed to promote inductive language learning, as well as to allow the student to acquire more autonomy and control of their own learning process. This active role is described by Johns as the learner acting as a "linguistic researcher" (2000, 108) or "language detective" (1997, 101).

As we have seen in Chapter 2, corpus linguistics is precisely one of the fields of linguistic inquiry that pinpointed the prevalence of collocations in language, thus it is only natural for corpora to be used in teaching collocations to language learners. Modern corpora comprising a vast amount of authentic language data can constitute a complementary resource besides collocation dictionaries for learning about or verifying native-like lexical combinations. Lewis (2000b, 198) notes that some benefits of using concordances for teaching collocations are that target items are always contextualized and that learners can access a large number of examples of the same item quickly. According to Woolard (2000, 40–41), concordances obtained from corpora can provide richer co-textual information than dictionaries (see also Kilgarriff 2009, 5), leading to an efficient exploration of the collocates of a word, although learners need to be trained in the use of this resource. He emphasizes that concordancing can be effectively applied in the correction of learners' production, while it renders students more sensitive to detecting whether two words constitute a native-like combination. This is in line with the idea that working with corpora equips students with useful strategies for the permanent and autonomous development of their collocational competence (see e.g. Moreno Jaén 2008, 232).

B. Collocation production and collocation learning with corpora

Descriptions of specific activities using language corpora and concentrating on collocations are provided in e.g. Moreno Jaén (2008), who proposes a complete teaching unit that consists of four stages and is aimed at introducing EFL learners to techniques which may help them further their collocational competence autonomously. The first stage of the teaching unit introduces the concept of collocation and the concordancing technique. The second stage involves activities which serve to raise learners' awareness to the importance of identifying and noticing collocations in the input, including error correction, translation and reformulation exercises. The third stage consists of reviewing collocations encountered in the previous stages, while the final stage focuses on the productive use of collocations prompting learners to develop corpus-based strategies supporting successful language production, as well as introducing them to further learning tools. In the context of SFL, the DDL approach is much less widespread, nevertheless, both Higueras García (2006) and Álvarez Cavanillas (2008, 75–78) suggest activities for learning Spanish collocations. The latter author, for instance, proposes an exercise which serves to introduce the concept of collocation to learners through concordances for keywords coming from a reading exercise, using data from the *Corpus de referencia del español actual* (Real Academia Española n.d.).

As concordances are often indicated for autonomous language learning, especially in the case of collocations, authors not only propose learning activities, but test their use as reference tools in language production tasks. Essentially, data from corpora is used either during the production task proper, i.e. without feedback from the teacher concerning learner output, or during a posterior revision of the learners' text, in which errors have been identified and marked by the teacher. Autonomous concordancing in language production was studied by Landure and Boulton (2010) and Yoon (2008), observing its positive effects during the completion of translation and writing tasks, respectively. In comparison, Gaskell and Cobb (2004) found that although learners used concordance lines pre-selected by their teacher to correct writing errors, most of them were reluctant to autonomously consult corpus while carrying out a writing task.

The only study focusing specifically on the use of multiword expressions aided by autonomous corpus consultation during oral language production tasks I have knowledge about is Geluso and Yamaguchi (2014). These authors developed learning activities and applied them in a semester-long DDL-based course with the aim of increasing EFL learners' repertoire of formulaic sequences, as well as their ability to use them in conversation. One of the tasks involved autonomous concordancing in order to identify typical uses of words in a topic chosen by the learners, which they had to practice in conversation with a classmate, and with a native speaker outside the classroom. The conversations were recorded and the subsequent analysis carried out by native judges found that target expressions were used correctly in most cases. Thus, the authors concluded that learners were generally able to identify prefabricated expressions as well as their correct uses in the concordance task, and that they could also transfer these expressions to their own language production with considerable success. Nevertheless, it was also observed that students had some difficulty in introducing novel expressions into spontaneous conversation and using them in a pragmatically correct ways.

Revision of marked errors with the help of concordance lines was the subject of several studies. Just like studies on online production, although most of these did not focus on the use of multiword expressions or collocations specifically, their results give some indication concerning how successfully language learners can correct collocation errors when provided with feedback in the form of concordance lines. In Chambers and O'Sullivan's (2004) study eight postgraduate English native-speaker students of French were instructed in the use of concordance software, and subsequently asked to correct segments of their own essays marked-up for revision with the help of a small semi-specialized corpus comprised of texts which had a similar topic to that of student writings. Contrary to their expectations, the authors claim to have found that more modifications and corrections were made in the case of grammatical errors (gender and agreement, prepositions and verb forms) than in the case of lexico-grammatical errors, including L1 interference errors and the use of an incorrect verb in verb+noun combinations. In contrast, in a subsequent study using the same methodology with undergraduate and masters students, O'Sullivan and Chambers (2006) found a similar rate of positive changes, i.e. corrections, made as compared to all correction attempts in the case of grammatical and lexical errors, the latter involving mostly cases of incorrect word choice or inappropriate vocabulary.

Tono et al. (2014) examined the success-rate in the revision of one marked segment each in short essays produced by 68 upper-intermediate and 25 lower-intermediate EFL students. They found that omission and addition errors were more successfully (80-100%) corrected as compared to misformation errors (62.8-75%). The authors hypothesized that this difference was due to the fact that some errors were easier to correct because the search terms to be introduced in the concordance program could be identified more straightforwardly. In the case of omission errors, for instance, the missing preposition in **talk* [*about*] *everything with her* could be easily supplied from concordances, in the same way as the correction of misformation errors was also more successful in cases of erroneous collocates as in **took a prize* (instead of *won a prize*).

A study focusing more explicitly on collocation errors is that of Wu, Witten and Franken (2010). These authors carried out a study to evaluate *FLAX,* a web-derived corpus and digital library software aimed at enhancing students' collocation use, (for more detail see Section 3.4.2.3) and to document the types of errors learners were able to correct using the tool. The nine EFL students who participated in the evaluation protocol were asked to complete an IELTS argument writing task selected by their teacher, which was marked for errors. Subsequently, participants were provided with the *FLAX* user guide and asked to revise their essays with the help of the tool. Errors marked in learner essays fell into two main categories: 1) grammatical errors, such as incorrect uses of verb forms and prepositions, misused plurals and articles, and missing verbs; and 2) lexical errors, consisting of wrong or inappropriate adjective+noun, verb+noun, noun+verb, etc. combinations. It is important to note that the tool being tested is different from traditional concordancers in that – as it is explained in the following subsection – searches are carried out in an n-gram library and results are organized according to the frequency of recurring patterns. It was found that 67% of all changes made by the students were successful, with the success rate being highest (100%) in the case of verb+preposition errors. Lexical errors in noun+*of*+noun and adj+noun combinations were also corrected successfully at a high rate, while verb+noun combinations proved to be the most problematic. The authors attributed this to the difficulty of selecting the verb conveying the intended meaning, given that, participants were often found to opt for a verb resulting more familiar, instead of the target verb, regardless of the context. Finally, learners were found more successful in correcting lexical errors, i.e. incorrect combinations of lexical items, than grammatical errors.

Finally, Frankenberg-Garcia (2012) in a study comparing the usefulness of dictionary definitions and corpus examples found corpus examples to be more efficient in aiding language production, measured in an error correction task. Importantly this was found to be the case even in when dictionary definitions contained the information necessary to resolve the correction task in the same way to corpus examples specifically selected for this purpose. Furthermore, students who were provided multiple corpus examples, similarly to a concordancing task, performed better than those who had to work with a single example, regardless of the fact that all examples contained the necessary information. This suggests that concordances constitute an efficient tool for language production – or at least error correction, however, as the author notes, her results merely provide evidence regarding that learners are able to make use of the corpus examples suitable for the task at hand, and naturally does not mean that they would be able to find them (p. 289).

In addition to learners' ability to search and interpret corpus data, some studies also provide empirical evidence on the effect of concordancing on retention. Cobb (1997) is one of the first empirical studies investigating this issue. In a complex experimental setup the author managed to confirm that the use of concordances enhanced retention of single vocabulary items, and concluded that corpus use may constitute a way to replicate the rich input of L1 acquisition in a FLT context. As part of their study already mentioned above, Gaskell & Cobb (2004) found that after having used concordance feedback for different types of writing errors, in a subsequent writing task, participants' error rate reduced significantly in the case of a number of error types for which feedback had been offered, such as word order, use of pronouns and punctuation.

Studies concerned with the retention of collocations in particular from concordancing activities include Chan and Liou (2005), Zaferanieh and Behrooznia (2011) and Huang (2014). Chan and Liou (2005) found that verb+noun collocations taught using an English-Chinese bilingual concordancer were retained at a higher rate than collocations taught through other methods as shown by the results of an immediate post-test taken after the treatment sessions. Nevertheless, the results of a delayed post-test suggested that many collocations were forgotten three and a half months after instruction, although, learners were more likely to remember combinations taught through concordancing, which suggests that this method can better assist students' long-term collocation learning. Zaferanieh and Behrooznia (2011) compared the effectiveness of web-based concordancing and traditional methods used in the instruction of Iranian EFL learners. Two groups of learners were taught collocations through the different instruction methods, and subsequently tested for target combinations. While the concordance group performed significantly better in the post-test on collocations which were not congruent with participants' L1, in the case of congruent collocations, no significant difference could be observed. Finally, Huang (2014) compared the effect of concordances and dictionary entries on the use of abstract nouns in combinations. She found a greater increase in the ratio of error-free use of target nouns in the case of the group of students who studied them in concordances than those who used dictionary entries. It was also observed that argumentative essays written by the corpus group contained a wider range of collocates. For instance, in the case of the noun *effect* the corpus group used the adjective collocates *harmful, terrible, serious, significant, negative, potential, numerous* and *positive,* while the dictionary group used combinations with the adjectives *bad, good, big* and *great.* In an essay written two weeks after the instruction, learners in the corpus group were found to have used more occurrences of the target nouns studied, while they also retained the combinations "quite accurately". Although the outcomes of Huang's (2014) experiment are rather promising, it is not clear whether the observed effects were merely induced by the medium of instruction or there were also differences between the methods and contents of materials used. The author does not provide detailed information on whether the dictionary entries examined by students contained a comparable amount of collocations to what was introduced in the concordances, furthermore, while it is mentioned that learners were explicitly instructed to study lexical combinations in the concordances, no information is given as to whether such activities were also done with dictionaries.

The results of the above mentioned studies are in general encouraging when it comes to learners' ability to interpret concordances as well as learning outcomes. What is more, questionnaires and/or interviews accompanying experimental studies often reveal positive attitudes of learners towards the concordancing method. In Huang's (2014) study, for instance, participants claimed that corpus data was helpful for learning collocations, grammatical patterns and memorizing the usage of words, while it also triggered incidental vocabulary learning. Nevertheless, learners also tend to refer to a number of drawbacks, such as the time consuming or tedious nature of working with concordance lines, the high amount of unknown words encountered, the lack of wider context (especially in the case of concordance print-outs), which may render interpretation of examples difficult, or the general difficulty of querying and analyzing corpus data (e.g. Geluso 2013; Huang 2014; Moreno Jaén 2008). Note that Frankenberg-Garcia (2012) found considerable individual differences in participants' performance, suggesting that students differ in their ability to interpret information found in corpus examples successfully. Accordingly, a number of authors (e.g. Yoon and Hirvela 2004) highlight the need to instruct language learners in the use of corpora and concordances. Still, Pérez-Paredes et. al. (2012) found, in a study observing learners' search strategies through logging their internet searches, that participants had problems in carrying out autonomous corpus searches even after instruction. The authors attributed this to that students, who can be considered "digital natives", are used to search engines like Google, equipped with underlying mechanisms for interpreting queries and enhancing search results, while, on the contrary, corpus tools are less advanced in that they require exact query terms to be introduced.

If it is the case that language learners are not able to carry out corpus queries satisfactorily, the question arises as to how to exploit concordance data for autonomous collocation learning effectively. The following subsection describes a number of tools – one of which has been mentioned above, as its use was tested in Wu, Witten and Franken (2010) – designed to facilitate learners' access to corpus data, and which can serve to enhance their collocational competence.

3.4.2.3 Online collocation learning tools

Two types of resources have been considered so far which language learners can use to find or study collocations: dictionaries and language corpora. An obvious drawback of the first results from size, i.e. the number of collocations included in a dictionary will be always limited, as well as that of the usage examples provided. In comparison, modern language corpora contain a vast amount of data, therefore they are likely to include occurrences of a larger amount of combinations. However, as it was discussed in the previous subsection, language learners are not necessarily able to satisfactorily exploit corpora, especially because they are likely to have problems carrying out queries. A way to overcome the difficulties posed by concordancing is to create tools that use corpus data, and are, at the same time, specifically designed for language learners. While – to my best knowledge – resources available for learners of Spanish are rather limited, EFL learners have at their disposal a somewhat wider selection of online resources which can aid their collocation production.

In what follows a series of such tools are described, including *FLAX*[17] (Witten et al. 2013), the *Sketch Engine*[18] (SkE, Kilgarriff et al. 2004; Kilgarriff et al. 2014), *Sketch Engine for Language Learners*[19] (SkELL, Kilgarriff et al. 2015), the automatically generated English[20] and Spanish[21] collocation dictionaries created with the word sketch technology (Kilgarriff et al. 2008), *StringNet*[22] (Wible and Tsao 2010), *Netspeak*[23] (Potthast, Trenkmann, and Stein 2010), *Collocation checker*[24] (Chang et al. 2008), *Collocation Inspector*[25] (J. C. Wu et al. 2010) and *Just the word*[26] (Edmonds n.d.), as well as *HARenEs*[27] (Alonso Ramos, García Salido, and Vincze 2014; Wanner, Verlinde, and Alonso Ramos 2013; Wanner et al. 2013). Most of these resources or their various components were designed to be used as reference tools, and can be classified according to the type of queries they allow and the feedback they offer. 1) Dictionary-like tools resemble conventional collocation dictionaries in that they allow to search for single words as well as in the way they display combinatorial information, 2) pattern-search tools offer search results in the form of n-grams or stings of words, while 3) collocation checkers allow users to verify whether a combination introduced is correct or not. *FLAX,* one of the above mentioned resources, not only constitutes a reference tool, but it also contains modules serving to engage users in a more 4) personalized learning experience, since it allows to create personal collocation lists and generate learning activities. The following sections describe these four types of tools or functions in order to provide an overview of existing collocation learning tools.

A. *Dictionary-like tools*

Dictionary-like tools work in an analogous way to electronic collocation dictionaries, i.e. they allow users to search for a single lexical item to find out what other

[17] http://flax.nzdl.org/greenstone3/flax?a=fp&sa=library

[18] http://www.sketchengine.co.uk

[19] https://skell.sketchengine.co.uk

[20] http://www.sketchengine.co.uk/ske.cgi?page=acd

[21] http://www.sketchengine.co.uk/ske.cgi?page=acd&article=a&language=Spanish

[22] http://www.lexchecker.org/

[23] http://www.netspeak.org/

[24] http://miscollocation-richtrf.rhcloud.com/

[25] http://inspector-richtrf.rhcloud.com/

[26] http://www.just-the-word.com/

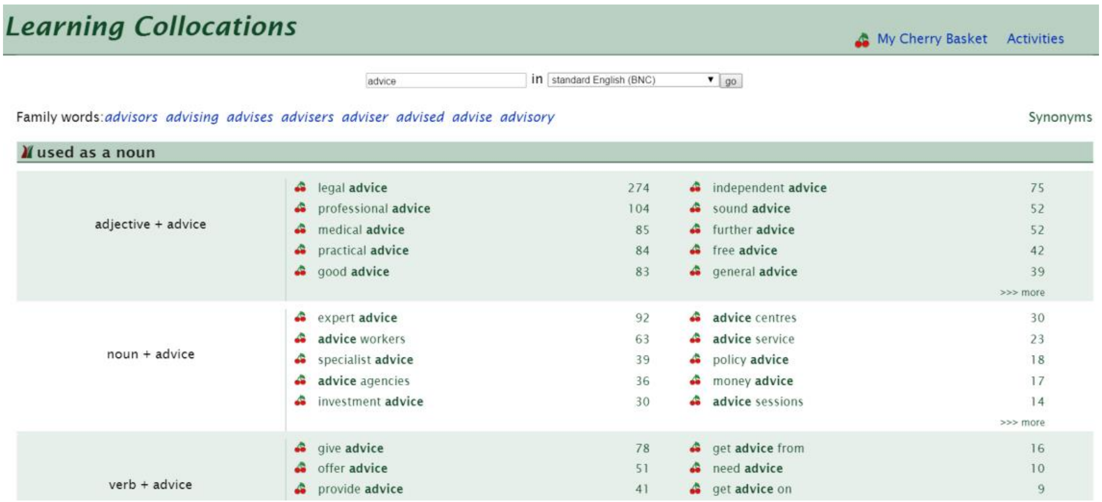
[27] http://harenes.taln.upf.edu/CakeHARenEs/

lexical units it is typically combined with. They differ from commercial dictionaries, however, in that their content is automatically generated on the basis of corpus data, and it is not edited by human lexicographers. Examples for such tools are the *Learning collocations* component of *FLAX*, the *Word Sketches* component of SkE, SkELL, the automatic Spanish and English language collocation dictionaries created from SkE data, the *Combinations* module of *Just the word* and the *Buscar colocaciones* (Collocation search) component of *HARenEs*.

The *Learning collocations* module of *FLAX* (Witten et al. 2013, 99–100) contains collocations from the *British National Corpus* (BNC), the *British Academic Written English Corpus* (BAWE) and the *Wikipedia*. The module was built through extracting segments corresponding to ten predetermined syntactic patterns[28] or their variants from each of the three corpora, matching them with combinations included in the *Web phrases* library (see below), and organizing them according to frequency. Collocations are thus identified through extracting strings of consecutive words which correspond to given sequences of parts of speech. For instance, for verb+noun collocations, the strings extracted can correspond to the following sequences: verb+noun (*make appointments*), verb+noun+noun (*cause liver damage*), verb+adjective+noun (*take annual leave*) or verb+preposition+noun (*result in a dismissal*).

The *Learning collocations* module can be used similarly to a collocation dictionary since when the user searches for a word in one of the three available corpora, the interface returns the most frequent combinations containing the target word organized into groups according to their syntactic pattern (see Figure 3). The authors highlight that the tool has several advantages over printed collocation dictionaries. One of these is full searchability, which means that users can find a given combination through searching for any of its elements, while it is also possible to search for the whole combination. The fact that in the search results collocations are displayed together with frequency information may help students prioritize learning (Wu, Franken, and Witten 2010).

[28] Syntactic patterns taken into account in the *FLAX Web Collocations* library are: 1) verb+noun, 2) noun+verb, 3) adjective(s)+noun(s), 4) noun+noun, 5) adverb+adjective, 6) adverb+verb, 7) noun+*of*+noun, 8) verb+adverb, 9) verb+adjective, 10) verb+*to*+verb.



Learning Collocations

My Cherry Basket Activities

advice in standard English (BNC) ▾ go

Family words: *advisors advising advises advisers adviser advised advise advisory*

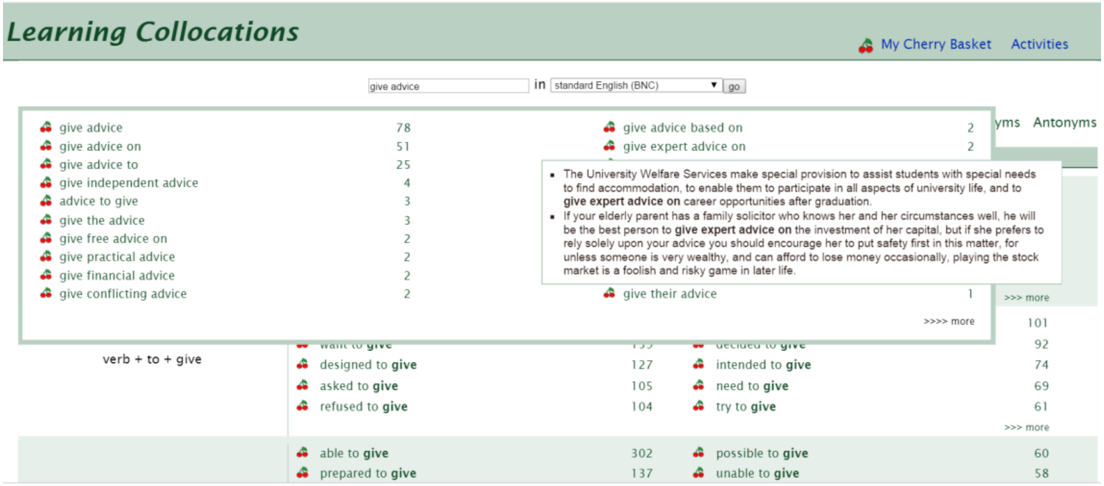
Synonyms

used as a noun

adjective + advice	legal **advice**	274	independent **advice**	75
	professional **advice**	104	sound **advice**	52
	medical **advice**	85	further **advice**	52
	practical **advice**	84	free **advice**	42
	good **advice**	83	general **advice**	39
				>>> more
noun + advice	expert **advice**	92	**advice** centres	30
	advice workers	63	**advice** service	23
	specialist **advice**	39	policy **advice**	18
	advice agencies	36	money **advice**	17
	investment **advice**	30	**advice** sessions	14
				>>> more
verb + advice	give **advice**	78	get **advice** from	16
	offer **advice**	51	need **advice**	10
	provide **advice**	41	get **advice** on	9

Figure 3 Screenshot of the search results obtained when searching for the noun "advice" in the BNC corpus using the Learning Collocation module of FLAX

Another important feature of the tool is that when clicking on a given collocation, or searching for a full combination, the user obtains a set of variants or patterns ordered according to frequency, which provide further information on the use of the combination (see Figure 4). As it is noted by the authors, learners can observe, for instance, whether nouns are generally used with an article (*make a difference* but not **make difference*) or whether they tend to be used in plural or singular (*make a decision/make decisions*, *make a living* but not **make livings*). Through clicking on a specific variant displayed, the user can access full sentence corpus examples, while the cherry icon to the right of each combination serves to store the expression in the user' personal dictionary, *My Cherry Basket*, see 3.4.2.3.D.



Learning Collocations

My Cherry Basket Activities

give advice in standard English (BNC) ▾ go

give advice	78	give advice based on	2
give advice on	51	give expert advice on	2
give advice to	25		
give independent advice	4		
advice to give	3		
give the advice	3		
give free advice on	2		
give practical advice	2		
give financial advice	2		
give conflicting advice	2	give their advice	1
			>>>> more

- The University Welfare Services make special provision to assist students with special needs to find accommodation, to enable them to participate in all aspects of university life, and to **give expert advice on** career opportunities after graduation.
- If your elderly parent has a family solicitor who knows her and her circumstances well, he will be the best person to **give expert advice on** the investment of her capital, but if she prefers to rely solely upon your advice you should encourage her to put safety first in this matter, for unless someone is very wealthy, and can afford to lose money occasionally, playing the stock market is a foolish and risky game in later life.

yms Antonyms

>>> more

				101
				92
verb + to + give	designed to **give**	127	intended to **give**	74
	asked to **give**	105	need to **give**	69
	refused to **give**	104	try to **give**	61
				>>> more
	able to **give**	302	possible to **give**	60
	prepared to **give**	137	unable to **give**	58

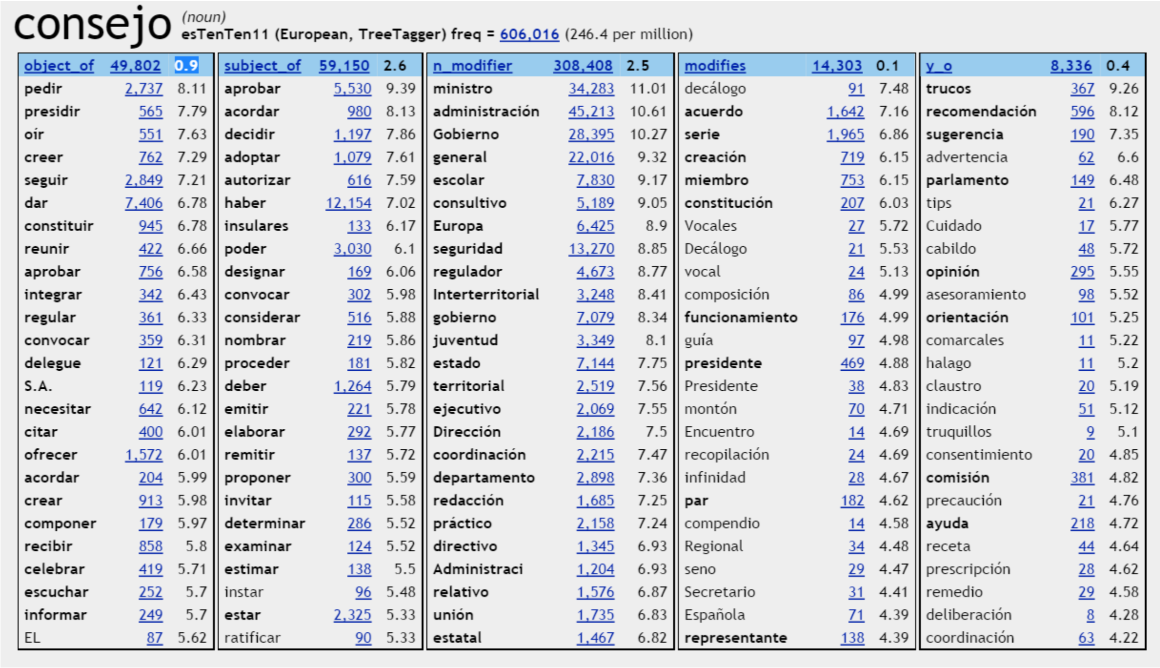
Figure 4 Variants of the collocation "give advice" and full sentence corpus examples shown in the Learning Collocation module of FLAX

The *Sketch Engine* (SkE) itself is not specifically aimed at language learners, given it was developed to be used by lexicographers and linguists. I have chosen to include it in this discussion since being a multilingual tool, it can be used to extract collocations from Spanish language corpora, and, in addition, its authors have also embarked on generating automatic collocation dictionaries, as well as creating a simpler and limited version of the query interface, which is more suitable for language learners (see below). The SkE is a complex corpus tool which incorporates multiple corpora in a great number of languages. One of its functionalities is constituted by the option of generating *Word Sketches*. These are described by the authors as "one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004, 105). Similarly to *FLAX,* the SkE uses shallow grammars based on strings of lexical units corresponding to specific sequences of parts of speech to extract collocations from corpora. These particular sequences are determined by Sketch Grammars. For instance, in the case of Spanish verb+noun combinations, corresponding strings are defined as shown in (18), i.e. a collocate verb (`VL.*`) of the noun may be preceded by a finite verb form (`V.fin`), and there may be an article (`ART`) and up to two elements corresponding to adjectives or adverbs (`ADJ|ADV`) between the collocate and the noun (`N`).

(18) =object/object_of

```
"V.fin"? 1:"VL.*" "ART"? "ADJ|ADV"{0,2} 2:"N.*"
```

When carrying out a query, the user is prompted to introduce a single lexical item and determine its part of speech. The tool returns lexical items which combine with the search term grouped according to the syntactic relationship established between the two, and ordered according to the logDice association score (Rychlý 2008), represented by a number in black (see Figure 5). When clicking on the number indicating raw frequency (the number in blue), the list of concordances containing the combination is displayed. As with the *Learning Collocations* module of *FLAX*, the *Word Sketches* function also has the advantage over paper collocation dictionaries of being fully searchable and containing a large amount of data, as well as providing access to a greater number of examples. In addition, as it was mentioned earlier, the SkE incorporates corpora in multiple languages, and can also be used with user generated corpora. Nevertheless, as mentioned above, the query interface, available for use through subscription, is rather complex and was not designed to be used by language learners.

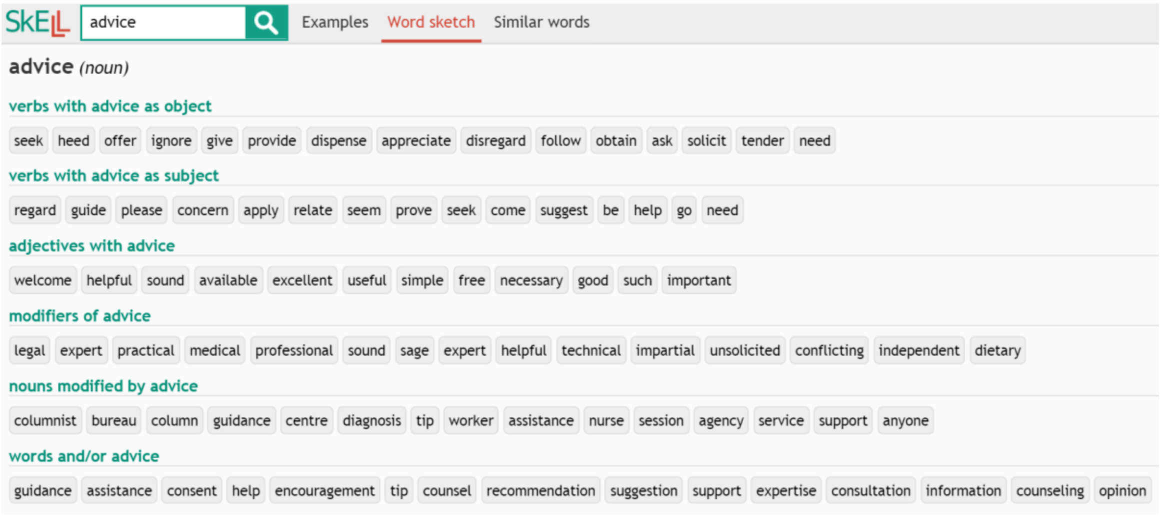


consejo (noun) esTenTen11 (European, TreeTagger) freq = 606,016 (246.4 per million)

object_of	49,802	0.9	subject_of	59,150	2.6	n_modifier	308,408	2.5	modifies	14,303	0.1	y_o	8,336	0.4
pedir	2,737	8.11	**aprobar**	5,530	9.39	**ministro**	34,283	11.01	decálogo	91	7.48	**trucos**	367	9.26
presidir	565	7.79	**acordar**	980	8.13	**administración**	45,213	10.61	**acuerdo**	1,642	7.16	**recomendación**	596	8.12
oír	551	7.63	**decidir**	1,197	7.86	**Gobierno**	28,395	10.27	**serie**	1,965	6.86	**sugerencia**	190	7.35
creer	762	7.29	**adoptar**	1,079	7.61	**general**	22,016	9.32	**creación**	719	6.15	advertencia	62	6.6
seguir	2,849	7.21	**autorizar**	616	7.59	**escolar**	7,830	9.17	**miembro**	753	6.15	**parlamento**	149	6.48
dar	7,406	6.78	**haber**	12,154	7.02	**consultivo**	5,189	9.05	**constitución**	207	6.03	tips	21	6.27
constituir	945	6.78	**insulares**	133	6.17	**Europa**	6,425	8.9	Vocales	27	5.72	Cuidado	17	5.77
reunir	422	6.66	**poder**	3,030	6.1	**seguridad**	13,270	8.85	Decálogo	21	5.53	cabildo	48	5.72
aprobar	756	6.58	**designar**	169	6.06	**regulador**	4,673	8.77	vocal	24	5.13	**opinión**	295	5.55
integrar	342	6.43	**convocar**	302	5.98	**Interterritorial**	3,248	8.41	composición	86	4.99	asesoramiento	98	5.52
regular	361	6.33	**considerar**	516	5.88	**gobierno**	7,079	8.34	**funcionamiento**	176	4.99	**orientación**	101	5.25
convocar	359	6.31	**nombrar**	219	5.86	**juventud**	3,349	8.1	guía	97	4.98	comarcales	11	5.22
delegue	121	6.29	**proceder**	181	5.82	**estado**	7,144	7.75	**presidente**	469	4.88	halago	11	5.2
S.A.	119	6.23	**deber**	1,264	5.79	**territorial**	2,519	7.56	Presidente	38	4.83	claustro	20	5.19
necesitar	642	6.12	**emitir**	221	5.78	**ejecutivo**	2,069	7.55	montón	70	4.71	indicación	51	5.12
citar	400	6.01	**elaborar**	292	5.77	**Dirección**	2,186	7.5	Encuentro	14	4.69	truquillos	9	5.1
ofrecer	1,572	6.01	**remitir**	137	5.72	**coordinación**	2,215	7.47	recopilación	24	4.69	consentimiento	20	4.85
acordar	204	5.99	**proponer**	300	5.59	**departamento**	2,898	7.36	infinidad	28	4.67	**comisión**	381	4.82
crear	913	5.98	**invitar**	115	5.58	**redacción**	1,685	7.25	**par**	182	4.62	precaución	21	4.76
componer	179	5.97	**determinar**	286	5.52	**práctico**	2,158	7.24	compendio	14	4.58	**ayuda**	218	4.72
recibir	858	5.8	**examinar**	124	5.52	**directivo**	1,345	6.93	Regional	34	4.48	receta	44	4.64
celebrar	419	5.71	**estimar**	138	5.5	**Administraci**	1,204	6.93	seno	29	4.47	prescripción	28	4.62
escuchar	252	5.7	instar	96	5.48	**relativo**	1,576	6.87	Secretario	31	4.41	remedio	29	4.58
informar	249	5.7	**estar**	2,325	5.33	**unión**	1,735	6.83	Española	71	4.39	deliberación	8	4.28
EL	87	5.62	ratificar	90	5.33	**estatal**	1,467	6.82	**representante**	138	4.39	coordinación	63	4.22

Figure 5 Word Sketch for Spanish noun "consejo" generated by the Sketch Engine

Sketch Engine for Language Learners (SKELL) is a simplified version of the original SkE interface. It allows obtaining *Word Sketches* presenting the search results in a format that resembles that of collocation dictionaries in that collocates are listed in groups corresponding to syntactic patterns (Figure 6). Through clicking on one of the lexical elements combining with the search term, the user can obtain a number of corpus examples. While this free access tool is admittedly more suitable for use by language learners than the original SkE, the information it provides is considerably more limited, while, it is only available in English.



advice (noun)

verbs with advice as object

seek heed offer ignore give provide dispense appreciate disregard follow obtain ask solicit tender need

verbs with advice as subject

regard guide please concern apply relate seem prove seek come suggest be help go need

adjectives with advice

welcome helpful sound available excellent useful simple free necessary good such important

modifiers of advice

legal expert practical medical professional sound sage expert helpful technical impartial unsolicited conflicting independent dietary

nouns modified by advice

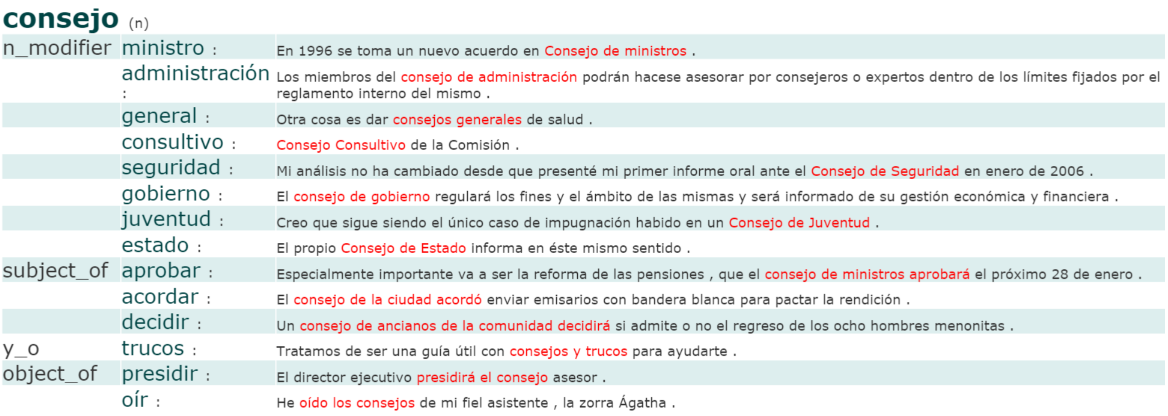
columnist bureau column guidance centre diagnosis tip worker assistance nurse session agency service support anyone

words and/or advice

guidance assistance consent help encouragement tip counsel recommendation suggestion support expertise consultation information counseling opinion

Figure 6 Word Sketch for the noun *advice* shown on the SkELL interface

The *word sketch* technology has also been exploited for the creation of automatic collocation dictionaries (Kilgarriff et al. 2008). Such dictionaries exist for English and Spanish, and render the information obtained from word sketches in a simplified format, through displaying a list of collocates, each accompanied by a full-sentence corpus example (see Figure 7). Similarly to the case of SKELL, although this tool is clearly more tailored to learners, data displayed is not nearly as rich as what is obtained with the original corpus tool. The user does not have access to more combinations, nor more examples than those shown on the results screen, furthermore, in contrast with paper collocation dictionaries, no semantic analysis of collocates is carried out to further group combinations.

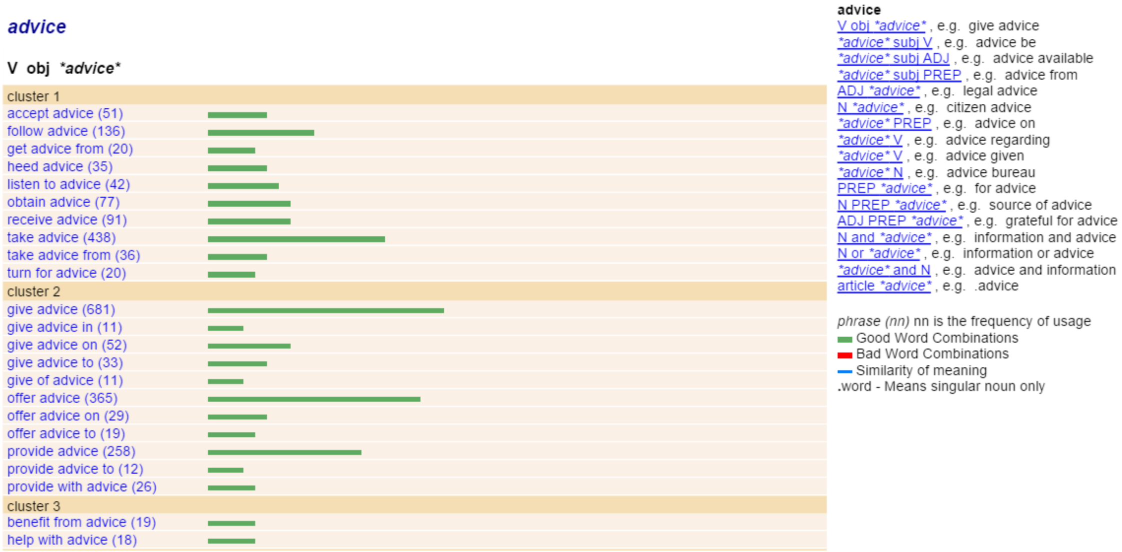


consejo (n)

n_modifier	ministro :	En 1996 se toma un nuevo acuerdo en Consejo de ministros .
	administración :	Los miembros del consejo de administración podrán hacese asesorar por consejeros o expertos dentro de los límites fijados por el reglamento interno del mismo .
	general :	Otra cosa es dar consejos generales de salud .
	consultivo :	Consejo Consultivo de la Comisión .
	seguridad :	Mi análisis no ha cambiado desde que presenté mi primer informe oral ante el Consejo de Seguridad en enero de 2006 .
	gobierno :	El consejo de gobierno regulará los fines y el ámbito de las mismas y será informado de su gestión económica y financiera .
	juventud :	Creo que sigue siendo el único caso de impugnación habido en un Consejo de Juventud .
	estado :	El propio Consejo de Estado informa en éste mismo sentido .
subject_of	aprobar :	Especialmente importante va a ser la reforma de las pensiones , que el consejo de ministros aprobará el próximo 28 de enero .
	acordar :	El consejo de la ciudad acordó enviar emisarios con bandera blanca para pactar la rendición .
	decidir :	Un consejo de ancianos de la comunidad decidirá si admite o no el regreso de los ocho hombres menonitas .
y_o	trucos :	Tratamos de ser una guía útil con consejos y trucos para ayudarte .
object_of	presidir :	El director ejecutivo presidirá el consejo asesor .
	oír :	He oído los consejos de mi fiel asistente , la zorra Ágatha .

Figure 7 Entry for the Spanish noun *consejo* 'advice' obtained from the automatic collocation dictionary derived from Sketch Engine data

A further free access tool which operates in a way analogous to an electronic collocation dictionary is the *Combinations* module of *Just the Word*. This tool – similarly to *FLAX* – is based on data obtained from the BNC. As it can be seen on Figure 8, when searching for a lexical item, the tool displays corresponding combinations grouped according to their syntactic pattern, while the length of the green bar to the right indicates association strength – technically, it indicates the t-score, representing the extent to which a combination occurs more often than it can be expected given the frequency of its component parts. The raw frequency of each combinations is shown by the number displayed in brackets next to it. Unlike in the case of any of the previous tools, here collocations are clustered into groups according to proximity of meaning, such that e.g. *accept, follow, get, obtain, receive* and *take an advice* constitute one group, and *give, offer* and *provide an advice* another. When clicking on a combination, the user can access corresponding concordance lines in the form of full sentences.



advice

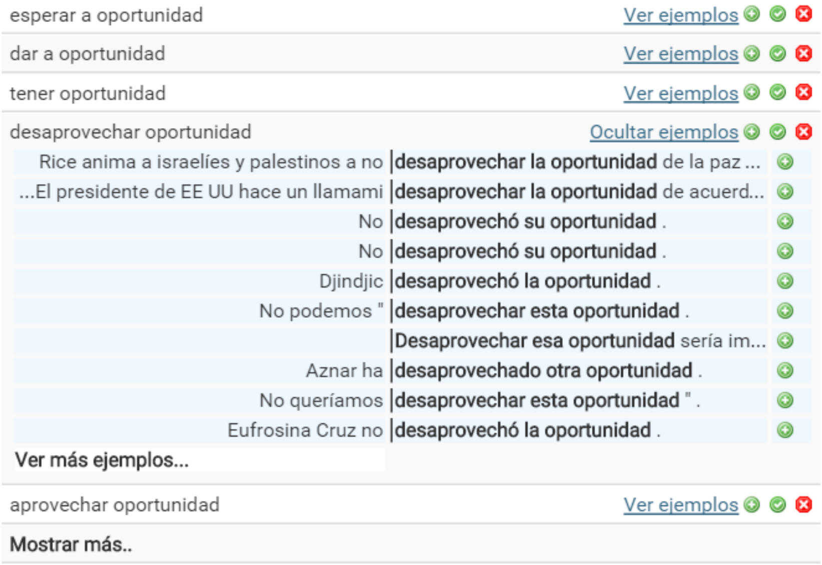
V obj *advice*

cluster 1
accept advice (51)
follow advice (136)
get advice from (20)
heed advice (35)
listen to advice (42)
obtain advice (77)
receive advice (91)
take advice (438)
take advice from (36)
turn for advice (20)
cluster 2
give advice (681)
give advice in (11)
give advice on (52)
give advice to (33)
give of advice (11)
offer advice (365)
offer advice on (29)
offer advice to (19)
provide advice (258)
provide advice to (12)
provide with advice (26)
cluster 3
benefit from advice (19)
help with advice (18)

advice
V obj *advice* , e.g. give advice
advice subj V , e.g. advice be
advice subj ADJ , e.g. advice available
advice subj PREP , e.g. advice from
ADJ *advice* , e.g. legal advice
N *advice* , e.g. citizen advice
advice PREP , e.g. advice on
advice V , e.g. advice regarding
advice V , e.g. advice given
advice N , e.g. advice bureau
PREP *advice* , e.g. for advice
N PREP *advice* , e.g. source of advice
ADJ PREP *advice* , e.g. grateful for advice
N and *advice* , e.g. information and advice
N or *advice* , e.g. information or advice
advice and N , e.g. advice and information
article *advice* , e.g. .advice

phrase (nn) nn is the frequency of usage
🟩 Good Word Combinations
🟥 Bad Word Combinations
— Similarity of meaning
.word - Means singular noun only

Figure 8 Combinations with the noun *advice* obtained with the *Just the Word*



esperar a oportunidad — Ver ejemplos
dar a oportunidad — Ver ejemplos
tener oportunidad — Ver ejemplos
desaprovechar oportunidad — Ocultar ejemplos

Rice anima a israelíes y palestinos a no |desaprovechar la oportunidad de la paz ...
...El presidente de EE UU hace un llamami |desaprovechar la oportunidad de acuerd...
No |desaprovechó su oportunidad .
No |desaprovechó su oportunidad .
Djindjic |desaprovechó la oportunidad .
No podemos " |desaprovechar esta oportunidad .
|Desaprovechar esa oportunidad sería im...
Aznar ha |desaprovechado otra oportunidad .
No queríamos |desaprovechar esta oportunidad " .
Eufrosina Cruz no |desaprovechó la oportunidad .
Ver más ejemplos...

aprovechar oportunidad — Ver ejemplos
Mostrar más..

Figure 9 Verbs combining with the noun *oportunidad* ‘opportunity’ obtained with the *Buscar colocaciones* component of *HARenEs*

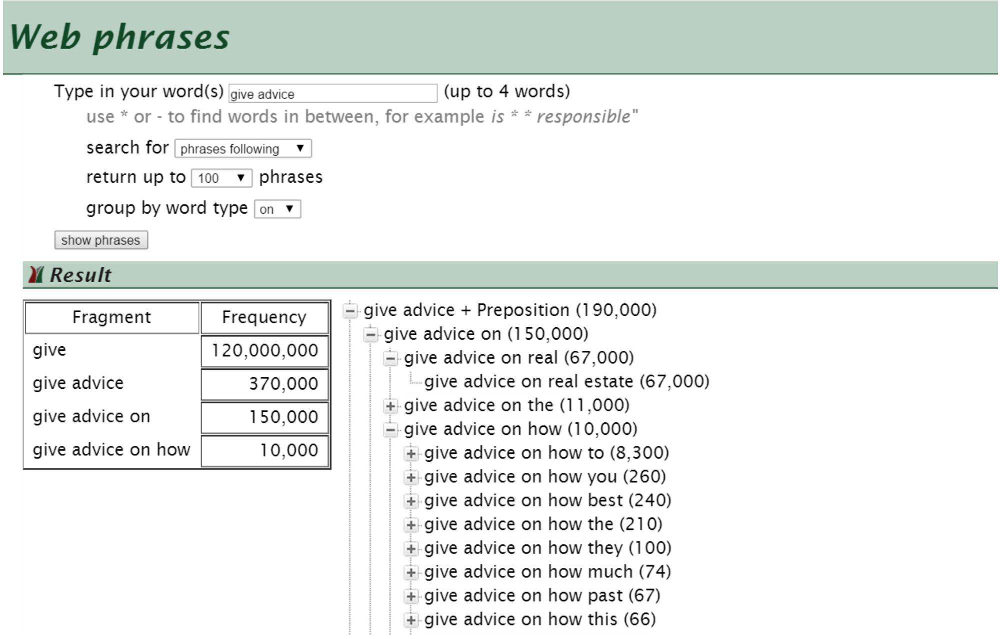
Finally, *HARenEs*, a collocation learning tool aimed at learners of Spanish also contains a module which allows users to obtain a list of co-occurring items once having introduced a lexical item as search term. The tool itself is being developed in the framework of a research project carried out by two research groups in collaboration, one at the University of A Coruña directed by Margarita Alonso Ramos, and another at the Pompeu Fabra University (Barcelona) directed by Leo Wanner. The component named *Buscar colocaciones* (Collocation search), similarly to the SkE, requires the user to introduce a word and chose its part of speech category, as well as that of the desired collocate. The results displayed consist of an expandable list of possible collocates, where corpus examples can also be consulted (see Figure 9). The combinations listed as a result of a query are extracted from a syntactically analyzed Spanish newspaper corpus, and are ordered according to a normalized Pointwise Mutual Information score (NPMI, Carlini, Codina-Filba, and Wanner 2014).

B. Pattern-search tools

As it was claimed above, string- or pattern-search tools offer search results in the form of n-grams or stings of words. N-grams are sequences of generally two to five consecutive words, in which corpora are often segmented for probabilistic analysis of language data. The output provided by pattern-search tools is thus different from that of dictionary-like tools reviewed previously, making them more suitable for searches concerning the use of specific collocations than for queries aiming to find the possible collocates of e.g. a given noun. Consequently, these tools can be used, for instance, to find out whether a certain verb requires a *to*+infinitive or a gerund complement, or what preposition is used with a given verb or noun, while they can also be used to search for items which can be used in a given context, i.e. to supply missing words during a production task. The following paragraphs describe in more detail the *Web phrases* component of *FLAX*, as well as the corpus search tools *Netspeak* and *StringNet*.

The *Web phrases* collection included in *FLAX* was developed using the Google n-gram corpus (Franz and Brants 2006). This corpus contains n-grams derived from publicly available English-language web pages, together with corresponding frequency information. The authors of *FLAX* processed this data base through intersecting it with the BNC vocabulary list in order to remove misspelled words, proper names, rare and other undesirable items to obtain the *Web phrases* collection, containing 14 million two-word phrases, and over 1300 million n-grams.

The query options available allow searching for words following or preceding a given item as well as supplying words that can occur between the elements of the search term, through the use of the wild card *. Thus, for instance, a search for "phrases following" *give advice* can be used to find the most common preposition co-occurring with this expression. As it is shown on Figure 10, word strings returned by the interface are organized according to the parts of speech of their components and their frequency, and they can be extended up to five words, after which access to corpus examples is provided. The effectiveness of this tool in correcting language learners writing errors was evaluated in Wu, Witten and Franken (2010), as discussed in Section 3.4.2.2.B.



Web phrases

Type in your word(s) give advice (up to 4 words)

use * or - to find words in between, for example *is * * responsible*"

search for phrases following ▼

return up to 100 ▼ phrases

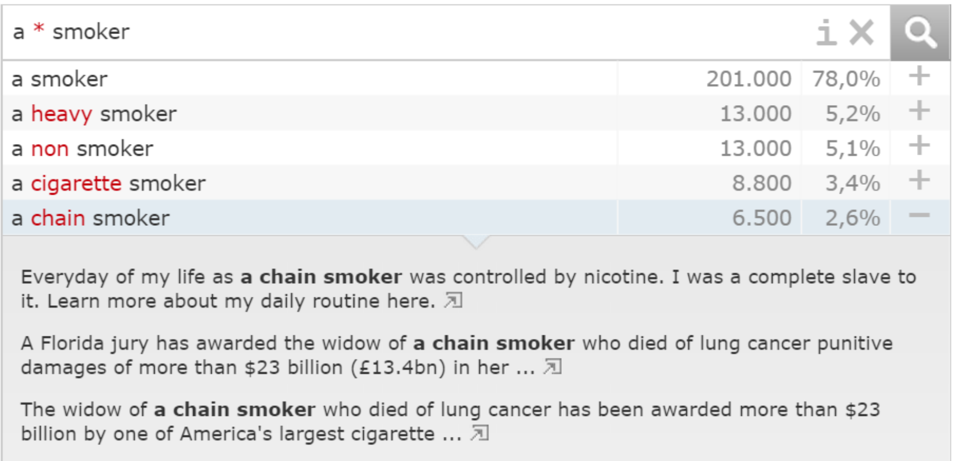
group by word type on ▼

Fragment	Frequency
give	120,000,000
give advice	370,000
give advice on	150,000
give advice on how	10,000

```
⊟ give advice + Preposition (190,000)
  ⊟ give advice on (150,000)
    ⊟ give advice on real (67,000)
        give advice on real estate (67,000)
    ⊞ give advice on the (11,000)
    ⊟ give advice on how (10,000)
      ⊞ give advice on how to (8,300)
      ⊞ give advice on how you (260)
      ⊞ give advice on how best (240)
      ⊞ give advice on how the (210)
      ⊞ give advice on how they (100)
      ⊞ give advice on how much (74)
      ⊞ give advice on how past (67)
      ⊞ give advice on how this (66)
```

Figure 10 Search results for phrases following *give advice* provided by the *Web Phrases* module of FLAX

Netspeak and *StringNet* are tools with similar functions. The first of these is based on the Google n-gram corpus, similarly to *FLAX*, and allows a number of wildcard searches, such as e.g. searching for one or more items embedded in a string, find the more frequent option of two or more expressions, or the most frequent word order in a given string (Potthast et al. 2010). For instance, as shown in Figure 11, a query for an embedded word occurring between the indefinite article *a* and the noun *smoker* returns expressions such as *a heavy smoker, a non smoker, a cigarette smoker, a chain smoker*. As this example demonstrates, the tool can serve to find a suitable adjective/collocate for a given context. Another type of query involving similar or synonymous expressions can be used to find combinations similar to the one introduced using the wildcard #. For instance, the search string *waiting for your #reply* allows the user to obtain a list of nouns that can be used in the same context, such as *response* and *answer*, as well as to verify the frequency of use of each. Note that, contrary to the case of *FLAX*, results here are ordered only according to frequency, are not and clustered according part-of-speech.



a * smoker

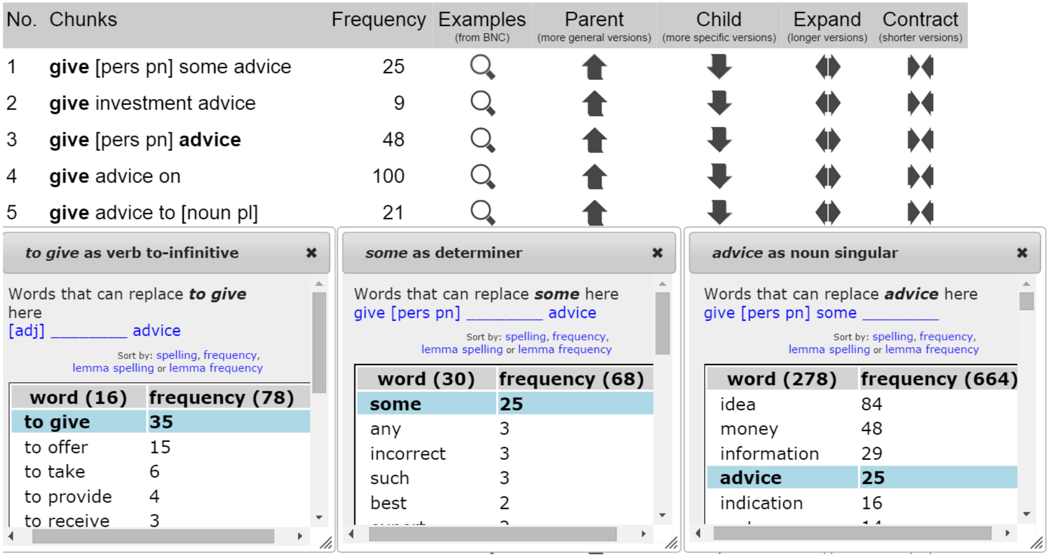
a smoker	201.000	78,0%
a heavy smoker	13.000	5,2%
a non smoker	13.000	5,1%
a cigarette smoker	8.800	3,4%
a chain smoker	6.500	2,6%

Everyday of my life as **a chain smoker** was controlled by nicotine. I was a complete slave to it. Learn more about my daily routine here.

A Florida jury has awarded the widow of **a chain smoker** who died of lung cancer punitive damages of more than $23 billion (£13.4bn) in her …

The widow of **a chain smoker** who died of lung cancer has been awarded more than $23 billion by one of America's largest cigarette …

Figure 11 Search results for the query for an embedded word occurring between the indefinite article and the noun *smoker* provided by *Netspeak*



No.	Chunks	Frequency	Examples (from BNC)	Parent (more general versions)	Child (more specific versions)	Expand (longer versions)	Contract (shorter versions)
1	**give** [pers pn] some advice	25	🔍	⬆	⬇	◀▶	▶◀
2	**give** investment advice	9	🔍	⬆	⬇	◀▶	▶◀
3	**give** [pers pn] **advice**	48	🔍	⬆	⬇	◀▶	▶◀
4	**give** advice on	100	🔍	⬆	⬇	◀▶	▶◀
5	**give** advice to [noun pl]	21	🔍	⬆	⬇	◀▶	▶◀

***to give* as verb to-infinitive**

Words that can replace ***to give*** here
[adj] _________ advice

Sort by: spelling, frequency, lemma spelling or lemma frequency

word (16)	frequency (78)
to give	**35**
to offer	15
to take	6
to provide	4
to receive	3

***some* as determiner**

Words that can replace ***some*** here
give [pers pn] _________ advice

Sort by: spelling, frequency, lemma spelling or lemma frequency

word (30)	frequency (68)
some	**25**
any	3
incorrect	3
such	3
best	2

***advice* as noun singular**

Words that can replace ***advice*** here
give [pers pn] some _________

Sort by: spelling, frequency, lemma spelling or lemma frequency

word (278)	frequency (664)
idea	84
money	48
information	29
advice	**25**
indication	16

Figure 12 Results of the query for the combination *give advice* offered by *StringNet*, with pop-up windows showing the most frequent items that can replace components *give*, *some* and *advice* in the string *give* [pers pn] *some advice*

StringNet was created from data coming from the BNC, and its database consists of what its authors refer to as "hybrid n-grams". Unlike traditional n-grams represented as strings of consecutive word forms or lemmas, hybrid n-grams consist of representations of the elements of the string on different levels, i.e. word form, lemma or part-of-speech. This type of representation allows to take into account the substitutability of each element through cross-indexing related n-grams (Wible and Tsao 2010, 25–27). For instance, in the case of the string *give him legal advice*, representing the word forms *him* and *legal* by their corresponding parts-of-speech allows establishing a relationship with similar strings, thus determining the substitutability of these items by other pronouns and adjectives, respectively.

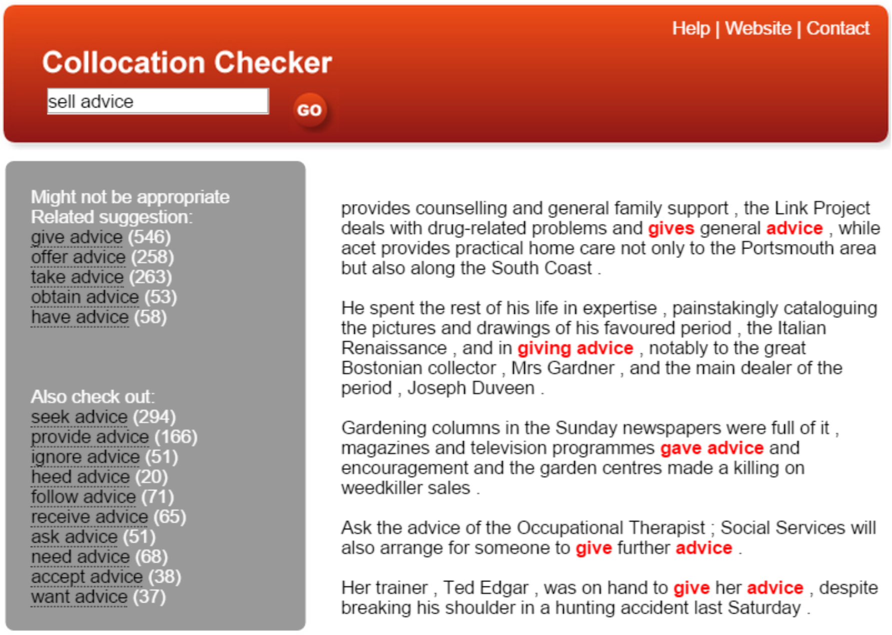
When the user introduces a lexical item or a string in the search box, the tool returns a list of hybrid n-grams ordered according to salience, determined by the MI score. As shown in Figure 12, the user is offered frequency information regarding each hybrid n-gram, together with the option to access corpus examples, related n-grams with more generic (parent) or more specific (child) structures, as well as extended and contracted versions of a given n-gram. Through clicking on any component of a string, a pop-up window is displayed which shows the most frequent words by which it can be substituted.

C. Collocation-checkers

Learning resources considered so far do not provide explicit feedback to learners regarding the correctness of the collocation searched for. This means that it is up to the user to interpret the corpus information and decide whether a certain combination constitutes a native-like expression, or it is necessary to find a more prototypical alternative to be used instead. Tools included in this section not only perform corpus searches, but also provide feedback to learners stating whether the collocation introduced in the query is correct or not. This judgment is based on frequency information and/or association measures through applying a threshold, similarly to how some collocation dictionary-like tools filter combinations to display.

The major difference in the format of the output constituted by the explicit feedback, however, may give the impression of dealing with an online tutor, instead of a reference tool. This is important to be borne in mind, since, naturally, the state of the art quality of automatic collocation error detection is far from reliable. Regardless of the limitations of current technology, as noted in Milton (2006, 132), users often appear to trust language checkers implicitly, while those recognizing their limitations believe that they are the result of bad engineering. In order to circumvent the imperfections of technology, Milton and Cheng (2010, 34) suggest exposing learners to authentic language and equipping them with strategies that enable contrasting their own production with target texts. This implies that, although collocation checkers provide straightforward feedback, in many cases, users are not fully exempt from interpreting the corpus data displayed. The tools described in this section include *Collocation Checker*, the *Alternatives* component of *Just the Word*, the *Comprobar colocaciones* module of *HARenEs*, and *Collocation Inspector*.

Collocation Checker is an online tool that allows the user to verify the correctness of verb+noun collocations, relying on a collocation library built from the BNC corpus (Chang et al. 2008). As shown in Figure 13, once the user introduces a query, *Collocation Checker* proposes a set of correction suggestions ordered according to relevance, together with frequency information and corpus examples. The first set of collocations labeled as “related suggestion” are meant to convey a similar meaning as that intended by the combination introduced by the user. The second set, marked as “also check out” contains other relevant verb+noun combinations containing the target noun. Importantly, the tool also offers usage examples and alternative combinations when the combination introduced by the user is judged to be correct.



sell advice

Might not be appropriate
Related suggestion:
give advice (546)
offer advice (258)
take advice (263)
obtain advice (53)
have advice (58)

Also check out:
seek advice (294)
provide advice (166)
ignore advice (51)
heed advice (20)
follow advice (71)
receive advice (65)
ask advice (51)
need advice (68)
accept advice (38)
want advice (37)

provides counselling and general family support , the Link Project deals with drug-related problems and **gives** general **advice** , while acet provides practical home care not only to the Portsmouth area but also along the South Coast .

He spent the rest of his life in expertise , painstakingly cataloguing the pictures and drawings of his favoured period , the Italian Renaissance , and in **giving advice** , notably to the great Bostonian collector , Mrs Gardner , and the main dealer of the period , Joseph Duveen .

Gardening columns in the Sunday newspapers were full of it , magazines and television programmes **gave advice** and encouragement and the garden centres made a killing on weedkiller sales .

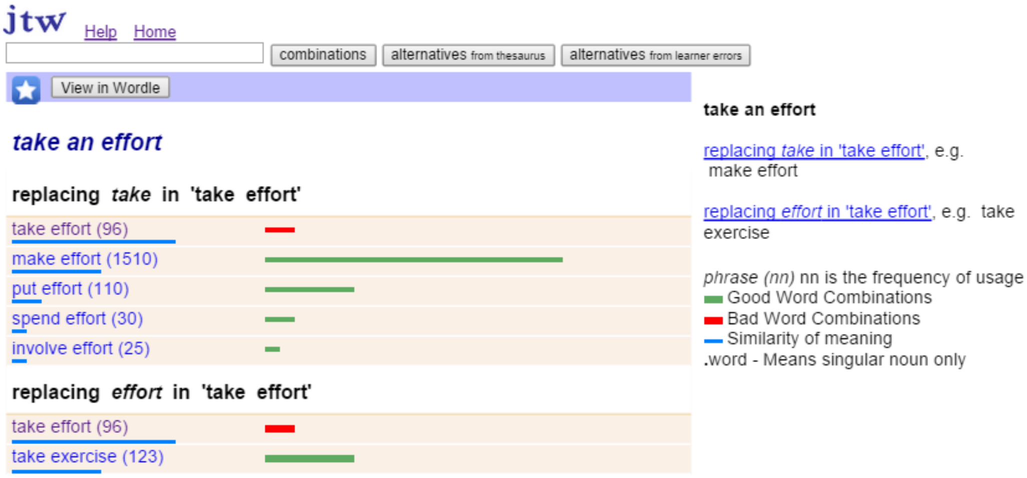
Ask the advice of the Occupational Therapist ; Social Services will also arrange for someone to **give** further **advice** .

Her trainer , Ted Edgar , was on hand to **give** her **advice** , despite breaking his shoulder in a hunting accident last Saturday .

Figure 13 Correction suggestions offered by Collocation Checker when introducing the combination "sell advice", with corpus examples for the suggested correct collocation “give advice” displayed on the right.

The drawbacks of the system include the fact that collocation verification is limited to verb+noun combinations, and that the query only accepts combinations in which the infinitive form of the verb is followed by a noun in the singular. This also applies in the case of collocations where the noun is obligatorily preceded by an article (*lay the table*) and combinations requiring the plural form of the noun (*take chances*), meaning that users are in fact asked to introduce an incorrect formulation of these combinations, i.e. without the article and with the noun in the singular form, respectively. Similarly, in the case of correction suggestions it is not overtly indicated whether the combination requires the use of a determiner or the plural form of the noun, thus the user has to deduce this information from the example sentences. Another limitation of the tool results from that it can only resolve cases where the error concerns the choice of the collocate verb. As we will see in the following chapter, language learners also make errors in the choice of the base. Nevertheless *Collocation Checker* cannot offer help to e.g. a Spanish speaking learner introducing the combination **give* [*a*] *conference* instead of *give* [*a*] *talk* or *give* [*a*] *presentation* resulting from the analogy with the Spanish noun *conferencia*, since the feedback displayed is limited to indicating alternative verbs that can be combined with the noun *conference*.

The *Alternatives* module of *Just the Word* serves to search for full combinations, and also offers information as to the correctness of the expressions queried. As illustrated in Figure 14, with the example **take an effort*, a red bar appearing to the right of the combination indicates that it is incorrect, while a green bar is used to mark correct combinations as well as suggested alternatives, with the length of the bars representing the degree of incorrectness or correctness, respectively. The frequency of each collocation is indicated by the number in brackets, while the length of the blue bar underneath each combination is aimed to represent its degree of semantic proximity to search term. Through clicking on a collocation, the user is provided access to corpus examples displayed in a full sentence format.



take an effort

replacing *take* in 'take effort'

take effort (96)
make effort (1510)
put effort (110)
spend effort (30)
involve effort (25)

replacing *effort* in 'take effort'

take effort (96)
take exercise (123)

take an effort

replacing *take* in 'take effort', e.g. make effort

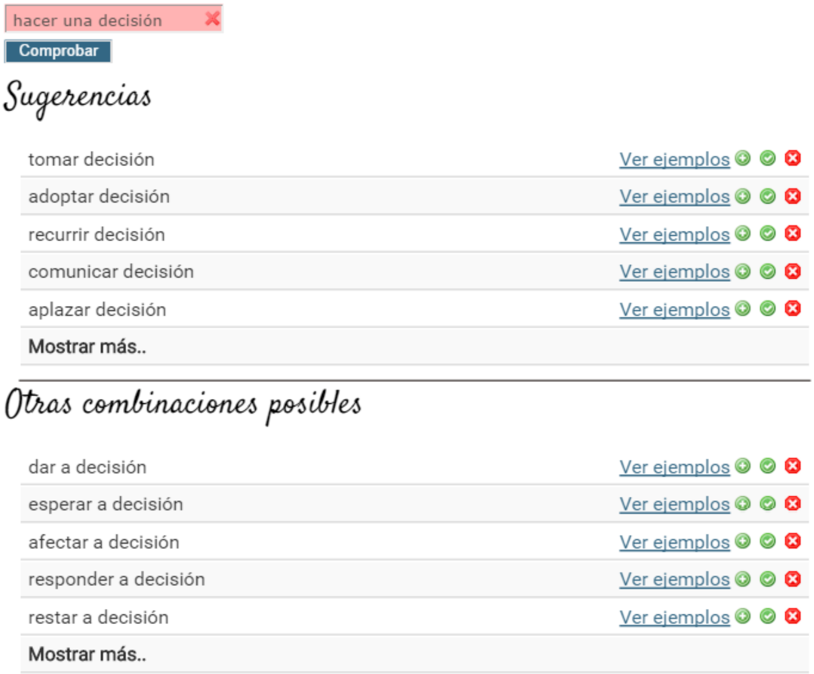
replacing *effort* in 'take effort', e.g. take exercise

phrase (nn) nn is the frequency of usage
🟩 Good Word Combinations
🟥 Bad Word Combinations
— Similarity of meaning
.word - Means singular noun only

Figure 14 Search results for the combination "take an effort" using the Alternatives module of Just the Word

This tool has a number of advantages over the *Collocation Checker*, described above. Firstly, it is not restricted to verb+noun combinations. Secondly, the queries admit search strings which contain a determinant and/or a preposition between a verb and a noun, plural nouns or different verb forms. Finally, as it can be seen in Figure 14, the interface not only offers suggestions for substituting the collocate (*take* → *make*, *put*, *spend*, *involve*), but also suggests alternatives for the base (*effort* → *exercise*).

The *Comprobar colocaciones* component of *HARenEs* allows users to verify word combinations in Spanish. As it is explained in Carlini et al. (2014, 6–7), the tool provides feedback regarding the acceptability of the collocation introduced by the user depending on whether it can be found in a reference corpus and whether its NPMI score (see 3.4.2.3.A), representing association strength reaches a threshold. Correction suggestions and alternative combinations corresponding to the syntactic pattern represented by the search term are retrieved from the corpus and ranked according to association strength[29].



hacer una decisión ✖

Comprobar

Sugerencias

tomar decisión — Ver ejemplos ⊕ ✓ ✖
adoptar decisión — Ver ejemplos ⊕ ✓ ✖
recurrir decisión — Ver ejemplos ⊕ ✓ ✖
comunicar decisión — Ver ejemplos ⊕ ✓ ✖
aplazar decisión — Ver ejemplos ⊕ ✓ ✖
Mostrar más..

Otras combinaciones posibles

dar a decisión — Ver ejemplos ⊕ ✓ ✖
esperar a decisión — Ver ejemplos ⊕ ✓ ✖
afectar a decisión — Ver ejemplos ⊕ ✓ ✖
responder a decisión — Ver ejemplos ⊕ ✓ ✖
restar a decisión — Ver ejemplos ⊕ ✓ ✖
Mostrar más..

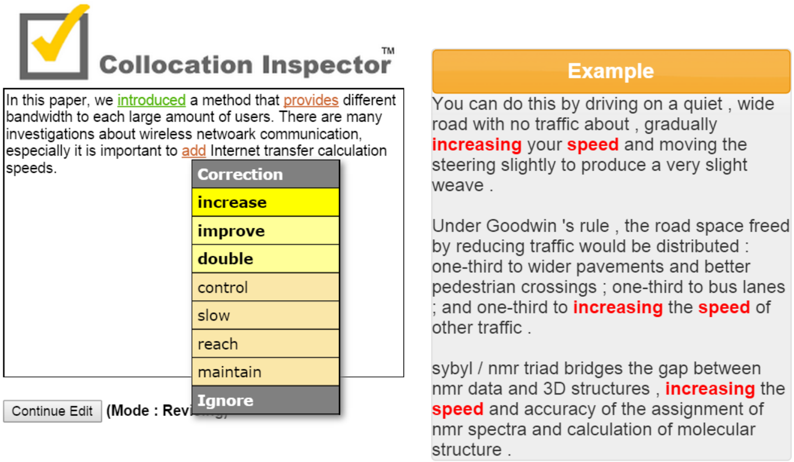
Figure 15 Feedback and suggestions offered by the *Comprobar colocaciones* component of *HARenEs* for the incorrect combination **tomar un esfuerzo* ‘take an effort’

As it can be seen on Figure 15, when a collocation is found incorrect, it is marked with the symbol ✖, and the search field is colored in red; on the contrary, collocations judged to be correct are marked in green and with the symbol ✔. Similarly to the case of

[29] Note that in the current stage, when the user introduced a combination with the patterns verb+object or verb+prepositional complement, the suggestions include combinations corresponding to both patterns displayed in two separate groups, see Figure 15.

Just the Word, suggestions or alternatives are offered both in the case of correct and incorrect combinations, while users have the option to expand the list of suggested combinations and to consult corpus examples.

The collocation checking function can be implemented as a writing aid tool which identifies collocations in a text introduced by the user, and verifies the correctness of these, suggesting possible improvements. One such tool is *Collocation Inspector*, proposed by J. C. Wu et al. (2010), while, the implementation of a – currently not available – writing aid functionality is also foreseen in the case of the *Comprobar colocaciones* component of *HARenEs* (cf. Alonso Ramos, García Salido, and Vincze 2014).



In this paper, we introduced a method that provides different bandwidth to each large amount of users. There are many investigations about wireless netwoark communication, especially it is important to add Internet transfer calculation speeds.

You can do this by driving on a quiet , wide road with no traffic about , gradually **increasing** your **speed** and moving the steering slightly to produce a very slight weave .

Under Goodwin 's rule , the road space freed by reducing traffic would be distributed : one-third to wider pavements and better pedestrian crossings ; one-third to bus lanes ; and one-third to **increasing** the **speed** of other traffic .

sybyl / nmr triad bridges the gap between nmr data and 3D structures , **increasing** the **speed** and accuracy of the assignment of nmr spectra and calculation of molecular structure .

Figure 16 Suggestions offered by the writing aid tool Collocation Inspector

Collocation Inspector identifies verb+noun collocations in the text introduced by a user, and offers alternatives or correction suggestions to replace the collocate in the case of correct combinations (with the verb marked in green, e.g. *introduce a method*) and incorrect collocations (with the verb marked in red, e.g. *provides different bandwidth*, *add Internet transfer calculation speeds*), respectively. When clicking on a verb in the list of suggested collocates, the tool displays a number of corpus examples, see Figure 16. The drawbacks of this tool include that, similarly to *Collocation Checker*, it is limited to suggesting alternatives for the collocate in verb+noun combinations, while, at its current stage of development, there is also much room for improvement both in the case of identifying combinations and offering suggestions. Nevertheless, *Collocation Inspector* does give an idea of the potential utility of collocation checking implemented as a writing aid tool.

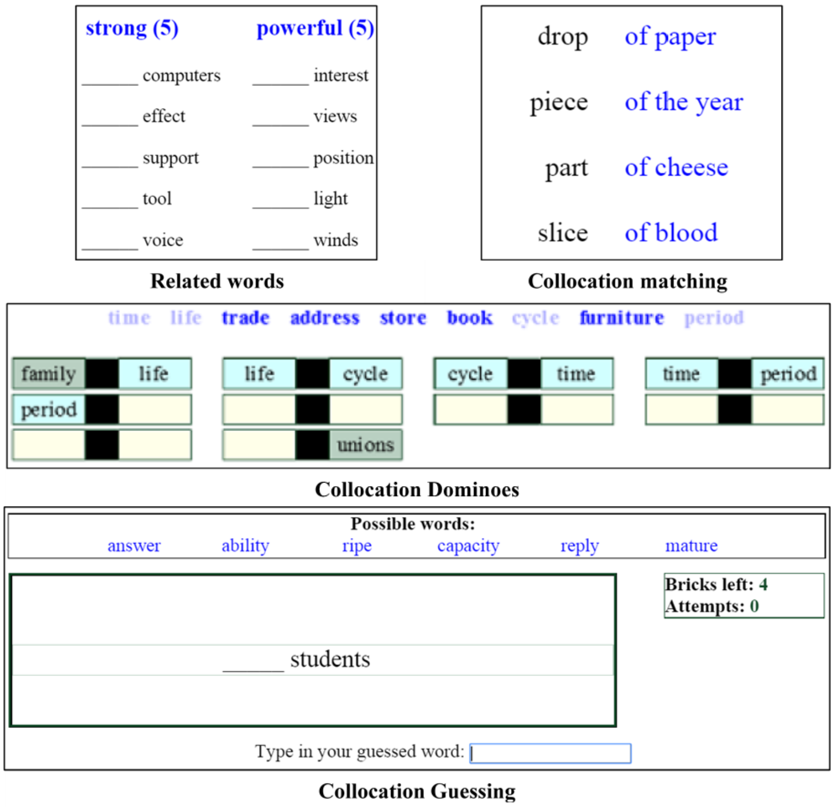
D. Personalized collocation learning

The previously reviewed functionalities of learning resources allow language learners to exploit corpus information as a reference tool. Nevertheless, the electronic interface also enables implementing modules providing a more personalized learning experience, as it was done in the case of the *FLAX* learning tool, which offers the option of generating collocation exercises, and includes a personal collocation dictionary component.

In addition to the *Learning collocations* and *Web phrases* components, *FLAX* also accommodates digital library collections containing texts suitable for learners of English with different proficiency levels, and offers the option of incorporating new texts, allowing, for instance, teachers to create learning activities for their students. The interface supports collocation learning through automatically highlighting and extracting combinations from the texts stored in the digital library or contributed by users. This means that texts from reading exercises become a resource for collocation learning. Besides, the site provides the option of further interacting with selected collocations through different activities.

The current version of the *FLAX* learning tool offers five different collocation learning activities[30]. Four of these consist of exercises presenting collocations as stand-alone items, i.e. without context (see Figure 17). The first two types of activities are aimed at helping to learn the distinction between commonly confused words through their collocations. In the activity *Related Words* the tool displays a list of items co-occurring with the target words selected (e.g. *strong* and *powerful*), and the users' task is to form correct combinations. *Collocation Matching* has a slightly different format, since here only one correct combination can be formed with each of the target items. The third activity type, *Collocation Dominoes*, consists of creating collocation chains where, similarly to the rules of dominoes, the adjacent boxes have to contain the same word. In the fourth type of activity, *Collocation Guessing,* combinations in which the same target word is missing are shown to the learner one by one until he or she manages to supply the missing item (Witten et al. 2013, 45–66; Wu 2010, 161–170).

[30] Wu (2010, 154–159) describes three more activities being *Common Alternatives*, where the user has to list different lexical items that can be combined with a target word, *Correcting Errors*, where the user has to substitute incorrect elements inserted in a text, and *Multiple Choice*. These activity types however are no longer available on the online interface.



Related words

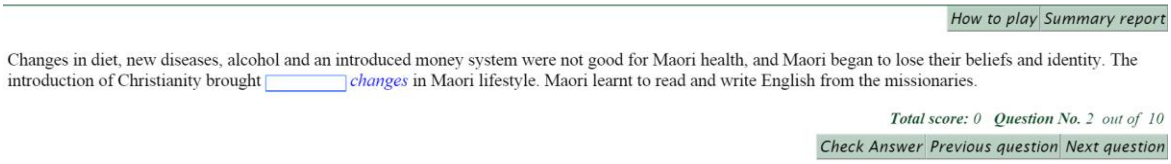
Collocation matching

Collocation Dominoes

Collocation Guessing

Figure 17 Collocation learning activities in *FLAX*

The fifth type of collocation activity offered by *FLAX* is *Completing Collocations*, consisting of a gap-fill exercise where a sentence containing a collocation is displayed with one of the members of the combination removed, see Figure 18. The sentences presented come from the texts included in the digital library collections.

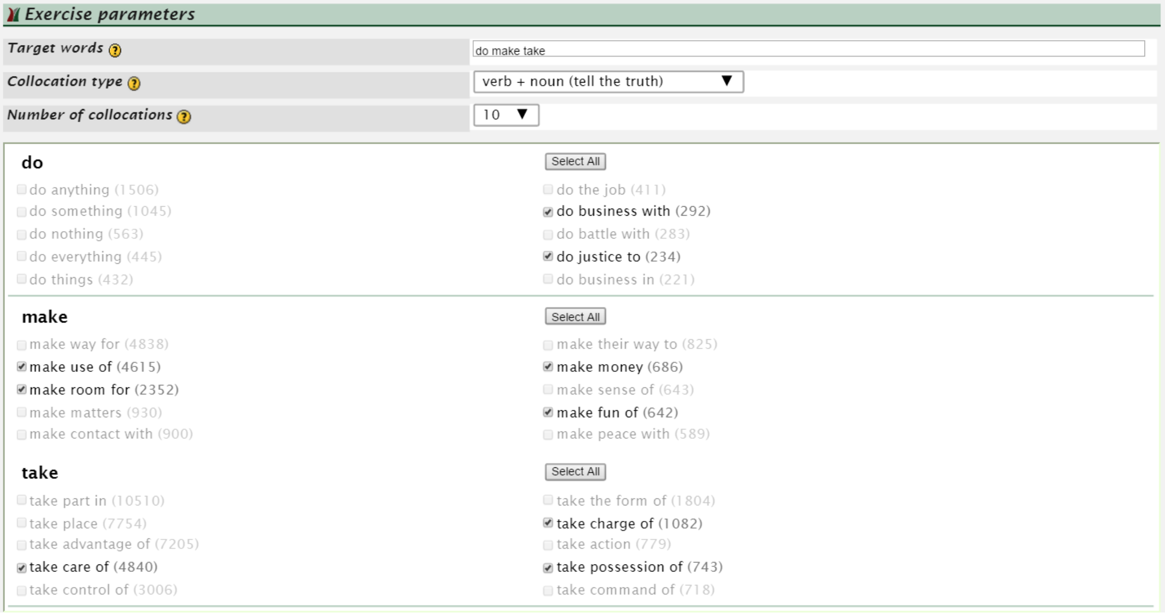


Changes in diet, new diseases, alcohol and an introduced money system were not good for Maori health, and Maori began to lose their beliefs and identity. The introduction of Christianity brought ______ *changes* in Maori lifestyle. Maori learnt to read and write English from the missionaries.

Total score: 0 Question No. 2 out of 10

Figure 18 The *Completing Collocations* activity type offered by *FLAX*

Importantly, users can generate practice activities according to their needs. For instance, as shown in Figure 19, to generate a *Matching Collocations* activity, one has to provide a list of target words to practice, chose a collocation pattern and determine the number of collocations to be retrieved per word. As a result, the interface lists frequent collocations corresponding to the query, from which the user can chose the target combinations to generate the learning activity.



Exercise parameters

Target words	do make take
Collocation type	verb + noun (tell the truth) ▼
Number of collocations	10 ▼

do

	Select All
□ do anything (1506)	□ do the job (411)
□ do something (1045)	☑ do business with (292)
□ do nothing (563)	□ do battle with (283)
□ do everything (445)	☑ do justice to (234)
□ do things (432)	□ do business in (221)

make

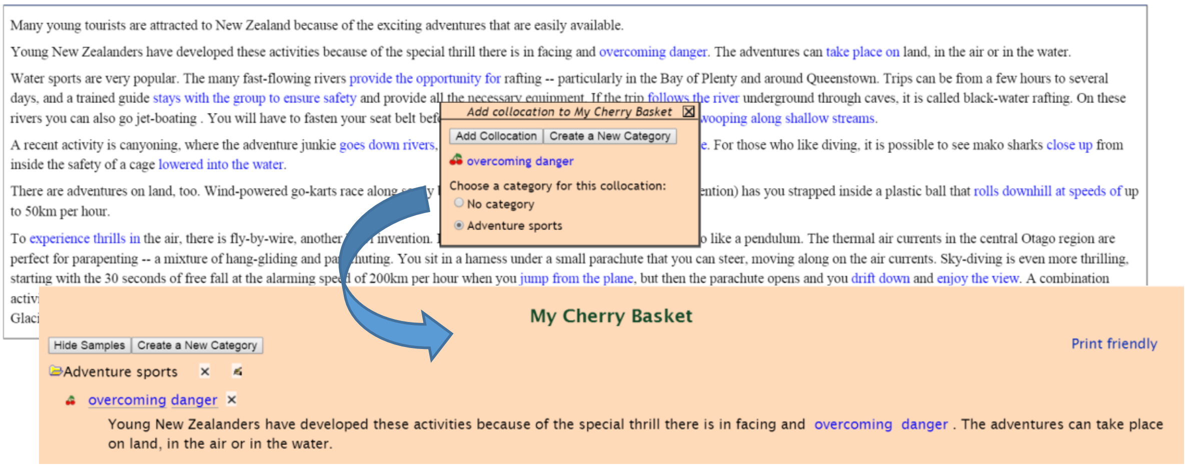
	Select All
□ make way for (4838)	□ make their way to (825)
☑ make use of (4615)	☑ make money (686)
☑ make room for (2352)	□ make sense of (643)
□ make matters (930)	☑ make fun of (642)
□ make contact with (900)	□ make peace with (589)

take

	Select All
□ take part in (10510)	□ take the form of (1804)
□ take place (7754)	☑ take charge of (1082)
□ take advantage of (7205)	□ take action (779)
☑ take care of (4840)	☑ take possession of (743)
□ take control of (3006)	□ take command of (718)

Figure 19 Screen allowing the user to generate *Collocation Matching* activities in *FLAX*

Finally, the *cherry-picking* functionality of *FLAX* allows the learner to "pick" any collocation and place it in their personal *cherry basket* (Wu 2010, 149–150). In order to save a combination while navigating the learning tool, the user has to click on the cherry icon (🍒) appearing next to the selected combination. As it is shown in Figure 20, each collocation selected by the user is saved together with the context it was found in. In addition, it is also possible to organize the content of the virtual vocabulary notebook through creating categories to sort collocations.



Many young tourists are attracted to New Zealand because of the exciting adventures that are easily available.

Young New Zealanders have developed these activities because of the special thrill there is in facing and overcoming danger. The adventures can take place on land, in the air or in the water.

Water sports are very popular. The many fast-flowing rivers provide the opportunity for rafting -- particularly in the Bay of Plenty and around Queenstown. Trips can be from a few hours to several days, and a trained guide stays with the group to ensure safety and provide all the necessary equipment. If the trip follows the river underground through caves, it is called black-water rafting. On these rivers you can also go jet-boating . You will have to fasten your seat belt bef… …wooping along shallow streams.

A recent activity is canyoning, where the adventure junkie goes down rivers, … …e. For those who like diving, it is possible to see mako sharks close up from inside the safety of a cage lowered into the water.

There are adventures on land, too. Wind-powered go-karts race along … …ention) has you strapped inside a plastic ball that rolls downhill at speeds of up to 50km per hour.

To experience thrills in the air, there is fly-by-wire, another … invention. … …o like a pendulum. The thermal air currents in the central Otago region are perfect for parapenting -- a mixture of hang-gliding and pa…huting. You sit in a harness under a small parachute that you can steer, moving along on the air currents. Sky-diving is even more thrilling, starting with the 30 seconds of free fall at the alarming spe…d of 200km per hour when you jump from the plane, but then the parachute opens and you drift down and enjoy the view. A combination activ… Glaci…

Add collocation to My Cherry Basket ☒

🍒 overcoming danger

Choose a category for this collocation:
○ No category
● Adventure sports

My Cherry Basket

Adventure sports ✕

🍒 overcoming danger ✕

Young New Zealanders have developed these activities because of the special thrill there is in facing and overcoming danger . The adventures can take place on land, in the air or in the water.

Figure 20 *Cherry-picking* a collocation in the *FLAX* learning tool

E. Summary of the characteristics of collocation learning tools

The previous subsections described a number of corpus-based online tools and components of more complex tools. As it was explained, the majority of these are designed to be used by language learners as reference resources aiding their use of collocations when completing L2 production tasks, while in the case of *FLAX*, corpus information is also exploited for the creation of learning activities. Reference tools were classified in three categories, including dictionary-like, pattern-search and collocation checker tools, while the personalized learning resources incorporated in *FLAX* were described in an individual section. Table 10 summarizes the characteristics of the collocation learning resources reviewed above. On a final note, it should be added that, although these tools have the potential of constituting valuable resources in the hands of a language learner, to my knowledge, their effectiveness – with the exception of *FLAX* (see Section 3.4.2.3) – has not been attested by experimental studies.

Tool	Language(s)	Component	Type	Collocation types	Query types	Type of feedback
FLAX	English	Learning collocations	Dictionary-like	multiple	Base, collocate whole combination	List of collocations organized according to syntactic pattern; frequency information; related n-grams; corpus examples
		Web Phrases	Pattern-search	multiple	1-4 word string to find preceding, following or embedded words	Expandable n-grams (up to 5 words); frequency; corpus examples
		Digital Library and Collocation Activities	Personalized learning	multiple	-	-
		Cherry Basket	Personalized learning	multiple	-	-
Sketch Engine	Multilingual (includes Spanish)	Word Sketches	Dictionary-like	multiple	Base, collocate	Co-occurring lexical items organized according to syntactic pattern; frequency; association measure; corpus examples
SkELL	English	Word Sketches	Dictionary-like	multiple	Base, collocate	Co-occurring lexical items organized according to syntactic pattern; multiple corpus examples
SkE: automatic collocation dictionaries	Spanish, English	-	Dictionary-like	multiple	Base, collocate	Co-occurring lexical items organized according to syntactic pattern; one corpus example per combination
Just the Word	English	Combinations	Dictionary-like	multiple	Base, collocate	Collocations organized according to syntactic pattern and clustered according to meaning; frequency; association strength; corpus example

Tool	Language(s)	Component	Type	Collocation types	Query types	Type of feedback
		Alternatives	Collocation checker	multiple	Whole collocation	Correction judgment; alternatives that can replace the collocate or the base; frequency; closeness of meaning; corpus examples
Netspeak	English	-	Pattern-search	multiple	Word string with wild cards	Matching n-grams; frequency; corpus examples
StringNet	English	-	Pattern-search	multiple	Single word or word string	Hybrid n-grams; substitutable lexical items; parent, child, extended and contracted n-grams; frequency; corpus examples
Collocation Inspector	English	-	Collocation checker	Verb + noun	Text	Identification of verb – noun collocations; correction judgment; alternative verbs with related meaning; other alternative verbs
Collocation checker	English	-	Collocation checker	Verb + noun	Verb + noun combination	Correction judgment; alternative verbs with related meaning; other alternative verbs
HARenEs	Spanish	Buscar colocaciones	Dictionary-like	multiple	Base (+ PoS of collocate)	Co-occurring lexical items corresponding to given PoS; multiple corpus examples
		Comprobar colocaciones	Collocation checker	multiple	Whole collocation	Correction judgment; alternatives that can replace the collocate; corpus examples

Table 10 Summary of the main characteristics of collocation learning tools

3.4.3 Summary: Learning collocations

The last sections were aimed to provide an insight into several aspects of teaching collocations, from pedagogical proposals giving special importance to lexical combinations to different existing resources language learners have at their disposal. We have seen that in the 1990s a number of authors concerned with revisiting the language teaching syllabus in the field of EFL recognized the crucial role of lexis and its intertwined character with grammar. Among these, Lewis’ (1997; 1993) Lexical Approach was the one that addressed collocations in most detail, advocating that single lexical items had to be always introduced together with their most significant collocations. In the context of SFL, a number of authors adopted Lewis’ ideas and proposed teaching sequences focusing on collocations, while, importantly, the *Plan curricular del Instituto Cervantes* (Instituto Cervantes 2006) also emphasizes the importance of expressions in the SFL syllabus.

A common thread in pedagogical proposals is the emphasis on familiarizing students with the notion of collocation and equipping them with strategies of autonomous learning, which can be applied outside the classroom. In line with this, the majority of the section was dedicated to reviewing tools that can be exploited for autonomous collocation learning: dictionaries, language corpora and online learning tools. When it comes to dictionaries, it was noted that while MLDs probably constitute the most commonly used dictionary type by learners and contain certain amount of collocational information, specialized collocation dictionaries provide the most amount of detail in the description of lexical combinations. The way in which collocation dictionaries present and organize combinatorial information was discussed in detail, followed by an overview of the results of experimental studies focusing on dictionary use in collocation reception and production tasks, as well as the description of a proposal for an alternative way of classifying collocations in the dictionary, which, implemented on an electronic dictionary interface, would allow more dynamic access to combinatorial information.

Language corpora are often indicated for autonomous language learning, and, more specifically, collocation learning, since they allow learners to study and deduce patterns of use from authentic language data. Accordingly, I reviewed a number of experimental studies focusing on learners' interaction with corpora, showing encouraging results in the case of online language production, error correction tasks, as well as post-task retention. Nevertheless, studies generally also encounter a number of problems regarding the use of corpora, including the degree of difficulty learners have in carrying out specific queries, together with the overwhelming amount of unstructured data obtained. In order to circumvent these difficulties, a number of corpus-based tools, some focusing specifically on collocations, have been developed for learners. In reviewing these learning resources, I proposed to classify those having the function of reference tools into three main groups: dictionary-like tools, pattern-search tools and collocation checker tools. Additionally, I also described functionalities incorporated by learning tools that provide a more engaging personalized learning experience.

3.5 Summary

The present chapter provided an overview of the numerous theoretical and pedagogical approaches, as well as empirical studies related to the broad area of inquiry comprised by the relationship of multiword expressions, specifically collocations, and second/foreign language learning. Three major issues have been explored in more or less detail: Firstly the main rationales for teaching and learning collocations were considered. These include language acquisition theories maintaining that patterns of language are acquired through the analysis of multiword strings memorized as chunks, together with language models claiming that a considerable part of native-like language is made up of unanalyzed sequences, which, at the same time, contribute to fluent language production. Secondly, I explored language learners' collocational competence and collocation use through two types of data: results of studies using testing methodology and studies exploring learner corpora. In general lines, it has been found that language learners' collocation knowledge correlates with their general language proficiency, while their use of lexical combinations is often characterized by the overuse of high frequency, favorite expressions as well as a major amount of erroneous combinations resulting from L1 transfer. Finally, the third part of the chapter was concerned with how different pedagogical proposals, influenced by linguistic research, integrated collocations in the language teaching syllabus in the fields of EFL and SFL, as well as with exploring the types of learning resources language learners have at their disposal to enhance their collocational competence and monitor their collocation use.

Notably, a great majority of existing empirical studies dealt with multiword expressions and collocations in the context of EFL. In order to gain information on SFL learners' collocational competence and their use of learning resources, three objectives will be pursued in the following chapters. Firstly, SFL learners' collocation production will be analyzed in a learner corpus study, secondly, a usability test involving the only existing online Spanish collocation dictionary, DiCE, will described, and thirdly, an empirical study will explore the ability of SFL students to correct different types of collocation errors using concordance feedback.

Chapter 4. Spanish as a foreign language learners' collocation production

4.1 Introduction

The present thesis aims to explore SFL learners' needs when it comes to resources aiding the use of collocations. In this context, I consider that the starting point of empirical inquiry should necessarily be constituted by the exploration of the collocation use of the target learner group. As we have seen in the previous chapter, existing studies dealing with language learners' knowledge of collocations have concentrated almost exclusively on learners of English. While, in the case of Spanish, although proposals for teaching collocations do exist (Higueras García 2006; e.g. Ferrando Aramo 2009) , these are merely based on the assumption that the multiword expressions in question are generally problematic for L2 learners, with no solid empirical basis concerning the actual collocation knowledge and use of the learners of this language.

4.2 Aims of the study

Learner corpora and methods used in corpus linguistics constitute a valuable instrument for studying language learners' collocation knowledge for two main reasons. Firstly, corpora consisting of texts produced by learners allow exploring the aspect of collocations deemed to be especially problematic, that is, production (see e.g. Hausmann 1985, 1010; Lewis 2000a, 134–136). Secondly, corpora are comprised of learner language in the form of continuous discourse, therefore they allow to gain insight into less guided production, containing combinations learners use spontaneously, as opposed to tests designed to elicit specific collocations. In what follows, the aims pursued by the study described here will be introduced in the light of the findings of previous learner corpus studies focusing on collocations, already reviewed in the previous chapter (see 3.3.2).

Although language learners have been found to use a lower overall amount of collocations than native speakers in a number of studies (e.g. Altenberg and Granger 2001; Granger 1998), it has not been confirmed, whether this applies to combinations of all types, especially since most researchers dealt with a confined group of collocations, usually corresponding to a given type of syntactic pattern. In fact, Siyanova and Schmitt (2008) found no significant difference in the use of adjective+noun collocations between learners and native speakers, while certain types of collocations, in particular, frequent or L1 congruent combinations, have been observed to be overused by learners (see e.g. Granger 1998; Lorenz 1999; Nesselhauf 2005). Furthermore, since most studies have focused on EFL learners, there is no solid evidence, whether these findings can be extended to the collocation use of learners of other languages. Consequently, one of the aims of the present study is to provide data concerning both the overall amount of collocations and combinations corresponding to different collocation types, as defined by syntactic pattern, used by SFL learners in comparison to native peers. The results are also compared with those of previous studies in order to draw some conclusions with regard to the generalizability of the observed usage patterns.

As for the amount of incorrect combinations used, in the case of EFL corpora, Nesselhauf (2005) and Laufer and Waldman (2011) claim to have found about one third of the collocations studied to be erroneous. In the case of SFL learners, Uriel Domínguez (2014) reports a considerably lower error rate in the production of upper-intermediate students, although it should be noted that her study takes into account both grammatical and lexical collocations (see 2.2.2 and 2.3.1), while Pérez Serrano (2014) finds over 50% of collocations to be erroneous in beginner and elementary level texts. While, as mentioned above, studies dealing with EFL learners generally looked at a limited group of collocations, studies focusing on a broader range of combinations do not provide data concerning whether different types of collocations are affected by errors at the same rate. In order to get a clearer picture of what type of combinations seem to pose more problems for SFL learners, the second aim of the present study is to provide data regarding the global collocation error rate observed, as well as the amount of correct and incorrect combinations corresponding to different syntactic patterns.

Although most authors focusing on collocations in the field of foreign language teaching express their concerns regarding erroneous use of combinations, only a few researchers analyze learner collocation errors in more detail, providing descriptions of the nature of erroneous combinations, as well as hypotheses concerning the source of errors and production strategies. Authors who present explicit classifications of collocation errors differ in the types of errors taken into account, the types of collocations studied, as well as in the aspects of particular errors their description focuses on. When it comes to the types of errors taken into account, one major difference lies in whether grammatical errors affecting a combination are considered to be collocation errors. Yorio (1989) and Nesselhauf (2005) describe as collocation errors the use of the plural form of a noun where only the singular is possible due to restrictions posed by the combination (e.g. ***have less chances to find a job*), or the lack of an obligatory determiner in a combination (e.g. ***run risk of*). In contrast, Howarth (1996) focuses only on errors involving the choice of lexical items. Comparison between error typologies proposed by different authors is sometimes problematic given they differ in the types of targeted collocations. While Yorio (1989) extracted all "conventionalized expressions" from the texts he studied, Howarth (1996) and Nesselhauf (2005) focused exclusively on verb+noun collocations, and Lorenz (1999) studied modifier+adjective combinations. As for the criteria used to classify and describe errors, error types are sometimes described with reference to the source of the error or production strategies, such as Yorio's (1989) "mixed idioms" (e.g. **it always strikes the mind* instead of *to strike sb* or *cross sb's mind*) and Howarth's (1996) "overlapping collocations" (e.g. **attach a role*) and "blending" errors (e.g. *achieve tasks*), while other error types distinguished by authors are more descriptive or generic in nature, such as Yorio's (1989) "lexical choice errors" (e.g. **made a great job*). Lorenz (1999) uses a two dimensional classification which specifies the erroneous element and describes the nature of the error, while Nesselhauf's (2005) error categories use three dimensions of classification, specifying the erroneous element, the nature of the error and its source (see 3.3.2.2).

The study presented here involved the creation of a comprehensive error typology, which can be applied to collocations independently of their syntactic structure and, similarly to Nesselhauf's (2005) work, describes three different aspects of the error in separate categories. The first category concerns the location of the error, while the second and the third categories correspond to descriptive and explanatory error analysis respectively. This error typology is used to provide both a qualitative description of the nature of collocation errors and detailed quantitative data concerning the prevalence of each error type. Therefore, this study aims to explore what aspects of collocations pose problems for learners, and reflect on the implications of characteristic learner errors in the case of collocation teaching and the development of learning tools.

In sum, the research questions constituting the focus of the learner corpus study described in the remainder of this chapter can be formulated as follows:

1) To what extent do SFL learners use collocations corresponding to different syntactic patterns, and how does the amount of collocations used by learners compare to native speakers' collocation use?
2) What is the error rate observed in the case of SFL learners' collocation production, and how is it distributed across different collocation patterns? In other words, do collocations corresponding to a given collocation pattern appear to be more problematic as reflected in the amount of errors?
3) What are the most prevalent types of errors that can be observed in SFL learners' collocation use, and what are their implications as to collocation teaching and the development of learning tools?

4.3 Methodology

The following sections describe the methodology adopted in the study presented in this chapter. First, I introduce CEDEL2, the learner corpus whose data has been exploited; second, I describe in detail the annotation scheme including the error typology applied for corpus annotation and, third, I explain the procedure followed during the annotation of the corpus.

4.3.1 The corpus: CEDEL2

4.3.1.1 Description of CEDEL2

The *Corpus Escrito del Español L2* (CEDEL2, Lozano 2009; Lozano and Mendikoetxea 2013) was created at the Autonomous University of Madrid, and currently constitutes the largest freely available written corpus of L2 Spanish[31]. It is comprised of two main components: an L1 English-L2 Spanish learner corpus and a native subcorpus. The two subcorpora were compiled following the same methodology so that the data obtained allows investigating different linguistic phenomena through comparing learner interlanguages corresponding to different proficiency levels and contrasting learner language with native production.

Data comprising CEDEL2 has been collected since 2006 through an online survey[32] consisting of three parts. In the first part of the questionnaire, referring to

[31] The *Corpus de aprendices del español* (CAES) released in October 2014 – posterior to when corpus annotation for the present study was carried out – contains 573718 tokens (Instituto Cervantes 2014).

[32] It is possible to contribute to CEDEL2 via the following URL: http://www.uam.es/woslac/start.htm

learning background, participants are asked to provide personal data (age, sex, education), as well as linguistic details (native language, parents' language, home language, time spent in Spanish speaking countries, etc.), their Spanish L2 proficiency level and proficiency level in other foreign languages according to self-assessment. The second part of the questionnaire consists of an online *proficiency test* developed at the University of Wisconsin (University of Wisconsin 1998). In the third and final part, participants are asked to write an essay on one of the twelve topics proposed by the authors of the corpus. Essay topics were compiled on the basis of themes typically appearing in different text books used in SFL classrooms, so that they differ in difficulty, and elicit the use of different types of linguistic structures, such as different verbal aspects, tenses and a wide range of vocabulary. Proposed topics include *What is the region where you live like? Talk about a famous person, What did you do last year during the summer holidays?* and *Talk about the problem of terrorism*. In addition, participants also have to provide information regarding whether they wrote the essay in or outside classroom, whether they had done any research to learn about the topic prior to writing, or used any language resources, such as dictionaries, spell checkers, etc.

Since material for the corpus can be submitted through an online form at any time, the corpus content is constantly growing. According to the most recent data published regarding corpus size (see Lozano and Mendikoetxea 2013), as of March 2011, CEDEL2 had reached a size of around 750,000 words, 27% of which belong to the native subcorpus and the remaining 73% to the learner subcorpus. A total of 711 native speakers and 1729 SFL learners submitted texts to the corpus, constituting 29% and 71% of the total number of participants respectively. According to the authors, these ratios reveal a relative balance between the number of participants and words contributed to the two subcorpora. As for participant profiles, most of the native speaker contributors received university education, and most of them were from Spain, with a minor number of participants from other Spanish speaking countries and the US, while 77% of SFL learners contributing to the corpus were students at US secondary schools or universities.

4.3.1.2 The subcorpora used in the present study

Since the study presented here involved a rather time demanding manual annotation procedure, only a portion of the CEDEL2 corpus used. This section describes the composition of both the corresponding learner and native subcorpora in detail.

Collocations were annotated[33] in a total of one hundred learner essays, consisting of 46420 words. These were produced by learners who obtained a minimum score of 77% in the external proficiency test administered when collecting the corpus data (see above); the average score of these learners was 91% (SD=6). The one hundred learners, whose essays were annotated, include 68 women and 32 men, corresponding to a variety of age groups, ranging between 17 and 74 years of age, although 61% of the learners were between 18 and 25 years of age. All learners were native speakers of English, two of them stated they were bilinguals, having additionally Italian and Japanese as their native language. Ten participants claimed they used a language other than English (Italian, Korean, Japanese, Portuguese, Polish or Spanish) as at least one of their home languages. As for their experience with Spanish, 84% of the participants had been learning the language for 3-10 years at the time of their contribution to the corpus (see Figure 21 for more detail), and 93% had spent time, in many cases, longer than six-month periods, in a Spanish speaking country, mostly in Spain (see Figure 22 for more detail).

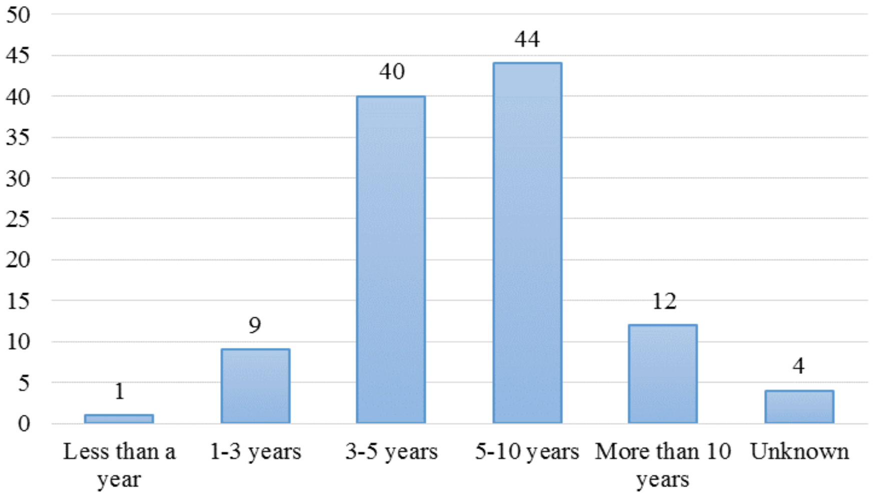


Figure 21 Time of study of SFL by participants contributing texts to the learner subcorpus used in the study

[33] The corpus annotation constituted part of a research project and was carried out through the collaboration of members of the research group, including the author of this thesis, lead by Margarita Alonso Ramos at the University of A Coruña.

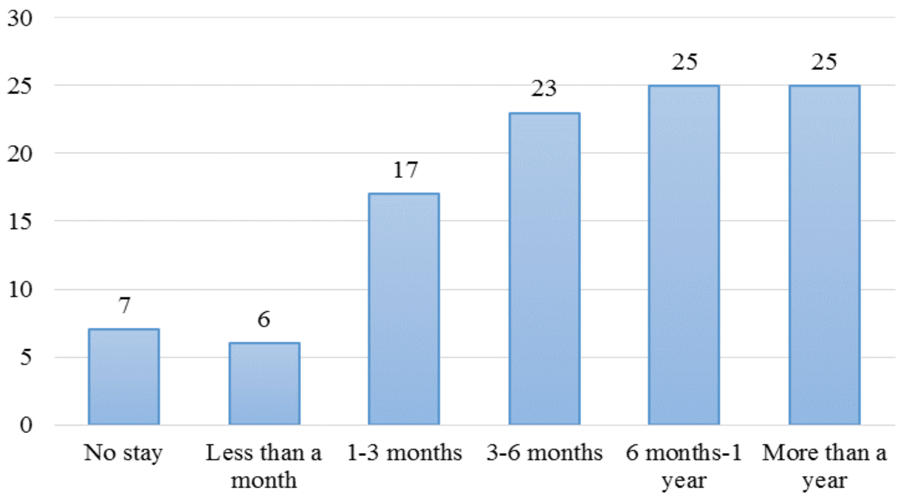


Figure 22 Length of stay in Spanish-speaking countries for participants contributing to the learner subcorpus used in the study

Finally, each of the twelve topics proposed by the designers of the corpus were represented in the learner subcorpus by at least one essay. As shown in Table 11, some of the essay topics were considerably more popular, while others were less well represented. As for resources used before or during the writing task, five of the total number of one hundred participants indicated that they carried out research on the topic prior to writing the essay, while ten participants used linguistic aids, mostly Spanish spellchecker; bilingual dictionary was used by two, thesaurus by one and native help also by one participant.

Essay topic (English translation)	**Number of learner essays**	**Number of native essays**
1. What is the region where you live like?	15	19
2. Write about a famous person	6	0
3. Summarize a movie you have seen recently	9	21
4. What did you do last year during your holiday?	15	5
5. What are your plans for the future?	10	11
6. Describe a trip you have been on recently	19	16
7. Write about an experience you have had	2	7
8. Write about the problem of terrorism in the world	1	3
9. What is your opinion about the new smoke-free law?	6	6
10. Do you think that gay couples should have the right to get married and adopt children?	10	10
11. Do you think marihuana should be legalized?	3	1
12. Analyze the main aspects of immigration	4	4

Table 11 Number of essays corresponding to each proposed topic included in the learner and native subcorpora used in the study

The native subcorpus used in this study comprises a total number of 103 texts and 29935 words. All native contributors were from Spain; 76 of them were women and 27 were men. Similarly to the case of learners, native participants represent a variety of age groups between 19 and 78 years of age, with 68% of participants between 25 and 40 years of age. As it can be seen in Table 11, except for essay topic 2, all other proposed topics are represented by at least one text in the dataset, however, similarly to the case of the learner subcorpus, the number of essays is not evenly distributed across the different topics.

4.3.2 Annotation scheme and error typology

When annotating the CEDEL2 subcorpora correct and incorrect collocations were manually identified and tagged for a number of attributes, such as the base, the collocate and the corresponding lexical function (LF). An important component of the annotation scheme was constituted by the collocation error typology developed through a preliminary exploration of the corpus data. The following sections describe in detail the error typology and the criteria applied during the annotation process, finally a brief introduction of *Knowtator* (Ogren 2006), the tool used to carry out the corpus annotation, is provided.

4.3.2.1 Collocation error typology

In order to study collocation errors, first it has to be established when to consider a collocation to be erroneous. As it is explained below, in Section 4.3.3.1, two sources were used when judging the correctness of a combination: native speaker intuition and data from the Spanish reference corpus CREA (Real Academia Española n.d.). It is also important to note that, besides the appropriateness of the choice of lexical items constituting the combination, similarly to Nesselhauf's (2005) work, aspects such as word order, use of articles, prepositions or particular word forms required for the correct formulation of the collocation were also taken into account.

The error typology presented here was meant to provide a systematic and detailed description of the nature of erroneous collocations produced by language learners, allowing for both qualitative and quantitative analysis. It was already noted that error typologies specific to collocation errors are rare, and a summary of how different types of errors were described in studies was provided in Section 3.3.2.2. As it was mentioned above, most studies limited their scope to a given type of combination, generally defined by a specific syntactic pattern, while, in contrast, the aim of the current study is to describe SFL learners' collocation production in a more comprehensive manner, taking into account a broader range of restricted lexical combinations. In order to do so, it was necessary to devise an error typology that would accommodate most types of combinations, and as many different types of errors as possible, for which the adoption of a multidimensional approach to error description, similar to the one used by Nesselhauf (2005), was found suitable.

Accordingly, the error typology proposed for the description of erroneous collocations produced by SFL learners represents three different aspects of the error in separate categories. The first, location dimension, specifies what element of the collocation is affected by the error. The second dimension models descriptive error analysis and distinguishes between three main categories, lexical, grammatical and register error. Finally, the third dimension represents explanatory error analysis. It concerns the source of the error, captured by the main categories of transfer errors, that is, errors reflecting L1 interference, and interlanguage errors, resulting from the incomplete knowledge of the L2, without L1 interference.

A. The location dimension

The location dimension captures whether the error concerns the _base_, the _collocate_ or the _collocation as a whole_ (see Table 12). It is often assumed that when producing a collocation, the primary difficulty is to choose an appropriate collocate that can be combined with a base. It can be expected, therefore, that language learners produce combinations with an erroneous collocate, such as for instance *_interrumpir una regla_ lit. ‘interrupt a rule’ instead of _romper una regla_ ‘break a rule’. However, learners may also choose a base incorrectly as in *_lograr un_ gol ‘achieve a goal (in sport)’ instead of _lograr un objetivo_ ‘achieve an aim’. In some cases both the collocate and the base are erroneous, e.g. *_pasar un testemuño_ ‘pass a testimony’ (from Portuguese _passar un testemuño_) instead of _dar testimonio_ ‘give testimony’, therefore, these were annotated as containing both a base and a collocate error.

Finally, some errors identified in the corpus can be described as affecting the collocation as a whole, instead of either the base or the collocate separately. This category includes, for instance, the use of incorrect collocation-like expressions to convey a meaning which should be expressed by a single word, as in *_poner apasionado_ ‘make passionate’ instead of _apasionar_ ‘to fascinate’, or incorrect uses of single-word forms instead of a collocation, such as *_misenterpretación_ 'misinterpretation' instead of _mala interpretación_ lit. ‘incorrect interpretation’.

LOCATION OF ERROR	Element of collocation	Collocate	**interrumpir una regla* lit. ‘interrupt a rule’ instead of *romper una regla* ‘break a rule’ **pasar un testemuño* ‘pass a testimony’ instead of *dar testimonio* ‘give testimony’
		Base	**lograr un gol* ‘achieve a goal (in sport)’ instead of *lograr un objetivo* ‘achieve an aim’ **pasar un testemuño* ‘pass a testimony’ instead of *dar testimonio* ‘give testimony’
	Whole collocation		**poner apasionado* ‘make passionate’ instead of *apasionar* ‘to fascinate’ **misenterpretación* 'misinterpretation' instead of *mala interpretación* lit. ‘incorrect interpretation’

Table 12 Classification of errors within the “location” dimension

B. The descriptive dimension

The descriptive dimension of the error typology is aimed at characterizing errors through contrasting the erroneous expressions to their target counterparts. Error categories describe the nature of errors through making reference to linguistic categories affected or to ways in which the surface structure of the target expression differs from that of the erroneous utterance. (see Ellis and Barkhuizen 2005, 60–61) The descriptive dimension is split into three main error categories: *lexical*, *grammatical* and *register errors*. The first two of these are divided into further, more specific error types. For a summary of all error types distinguished within the descriptive dimension see Table 13.

A total of five subtypes are distinguished within the category of lexical errors. The first two can be defined in general terms as referring to the erroneous choice of a lexical element, either the collocate or the base:

(1) In the case of *substitution* errors, the affected element of the collocation is substituted by an inadequate but existing L1 or L2 lexical item, such as in the case of **llenar un puesto* ‘fill a position’ instead of *cubrir un puesto.*

(2) On the contrary, in the case of *creation* errors, the erroneous form substituting the affected element of the combination is a non-existing lexical item, i.e. an item “created” by the learner. An example for this type of error is *tiene *limitades* ‘have limits’ where the non-existent Spanish word **limitad* is supplied by the learner instead of the correct form *límite*.

The remaining three types of lexical errors concern the collocation as a whole.

(3) The error type named *analysis* refers to cases already mentioned above, when the language learner uses a non-conventional combination resembling a collocation to convey a meaning which is correctly expressed by a single lexical unit in Spanish. For instance, the combination **hacer de cotilleo* lit. ‘to make gossip’ produced by a learner can be seen as an attempted support verb+noun combination, whereas the correct form to use would be the single verb *cotillear.*

(4) *Synthesis* errors consist of the use of a nonexistent single lexical item instead of a collocation, such as in the case of **misinterpretación,* mentioned above, or the use of the form *snowboard* as a verbal element, instead of the combination *hacer snowboard* ‘to do snowboarding’.

(5) The third type of error affecting the collocation as a whole involves cases when an otherwise *correct combination is used in an incorrect sense*, such as in (19) where, in order to express the correct meaning, the combination *aliviar el estrés* ‘ease the stress’ should be substituted for *aumentar el estrés* ‘increase the stress’. Note that, strictly speaking, the erroneous and the target expression only differ in the lexical item used as the collocate, nevertheless, the error is described as affecting the whole combination. This decision was made on the basis of the hypothesis that the learner may know that the two lexical items involved prototypically go together, i.e. constitute a collocation, but does not have sufficient knowledge of the meaning of the combination (see also Nesselhauf 2005, 167–168). Another example, shown in (20), involves the confusion of the expressions *no dar la gana* ‘can’t be bothered to do sg’ and *tener ganas* ‘want to, feel like doing sg’, with different syntactic configurations.

(19) *al oirlo hablar, tengo que apagar al aparato para no *aliviar el estrés* ‘when I hear him [George W. Bush], I have to turn off the television in order to not to ease the stress’

(20) *no *le da la gana de aprender español* ‘[speaking of immigrants] (he or she) can’t be bothered to learn Spanish’ instead of *no tiene ganas de aprender español* ‘(he or she) doesn’t want to learn Spanish’

Grammatical error types make reference to the linguistic category affected by the error. A total number of six subtypes of grammatical errors were identified in the corpus:

(1) *Determination* errors refer to missing or incorrect uses of a determiner when it constitutes a deviation from the correct collocation structure. For instance, in the combination *tener el derecho de* ‘have the right to’ the use of the definite article is obligatory, therefore **tienen derecho de* is considered to be a collocation error, similarly to **dimos bienvenidas* lit. ‘we gave welcomes’ instead of *dimos la bienvenida* lit. ‘we gave the welcome’, i.e. ‘we welcomed’. Note that, in order for the erroneous use of a determiner to qualify as a collocation error, it has to do with the collocation structure, therefore e.g. the omission of an article required by a modifying relative clause, as in (21), or the lack of number agreement in determiners were not annotated.

(21) **no le tiene respeto que merece* instead of *no le tiene el respeto que merece* 'she does not have the respect for him that he deserves'

(2) *Number* errors concern the use of a singular noun form when a plural noun is required in a collocation, or the use of a plural noun instead of a singular noun, as in **tienen prejuicio* instead of *tienen prejuicios* 'they have preconceptions'. Similarly to the case of determination errors, lack of number concordance was not taken into consideration.

(3) *Gender* errors usually surface as incorrect gender concordance of the forms of determiners or adjectives, as in *convertirse *al religión* 'convert to the religion' where the feminine gender noun *religión* is preceded by the masculine form of the article *el.* These cases may be interpreted as resulting either from the learner failing to apply the concordance rule or from their lack of familiarity with the grammatical gender of the nominal element. According to the first interpretation, gender errors would not be considered to constitute collocation errors, since, although they affect the combination, they result from the learners' imperfect knowledge or use of L2 grammar. In contrast, according to the second interpretation, the error results from the incomplete knowledge of a lexical item involved in the collocation, thus its inclusion in an error typology specific to collocations is justified. Since, based solely on learners' production, it cannot be unambiguously decided which of the two possible interpretations applies in each case, all gender errors were annotated, the default assumption being that the learner lacked sufficient knowledge of the gender of the nominal element found in the collocation.

(4) *Government* errors concern cases of missing or erroneous uses of elements specified in the government structure of the base or the collocate. For example, in the case of **hablando al teléfono* instead of *hablando por teléfono* the learner used an incorrect preposition, while in **no lo tiene respeto* the accusative pronoun *lo* is used instead of the dative pronoun *le.*

(5) Incorrect uses of Spanish pronominal verbs were tagged as *pronoun* errors. This type of verbs require the obligatory use of the reflexive pronoun, therefore the omission or insertion of an unnecessary pronominal element can be considered to be the result of the learner's insufficient knowledge of a verb constituting an element of the collocation. An example is the case of [*me*] **muero de ganas* ‘I’m dying to do sg’ where the pronoun was omitted.

(6) The last type of grammatical descriptive errors is *incorrect word order*. Although Spanish word order is generally variable, in the case of certain collocations, especially of the type noun+adjective, the elements of the collocation conventionally appear in a certain order. Deviations were found in e.g. **blanco vino* ‘white wine’ and **reputación mala* ‘bad reputation.

DESCRIPTIVE DIMENSION	**Lexical errors**	**Substitution**	**llenar un puesto de trabajo* ‘fill a position’ instead of *cubrir un puesto*
		Creation	**hacer un llamo* ‘receive a call’ instead of *hacer una llamada*
		Analysis	**hacer de cotilleo* lit. ‘to make gossip’ instead of *cotillear* ‘to gossip’
		Synthesis	**escaparatar* ‘to window shop’ instead of *ir de escaparates*
		Correct combination with incorrect meaning	**no le da la gana* ‘he/she can’t be bothered’ instead of *no tiene ganas* ‘he she doesn’t want to’
	Grammatical errors	**Determination**	**tienen derecho de* instead of *tienen el derecho de* ‘they have the right to’
		Number	**tienen prejuicio* instead of *tienen prejuicios* ‘they have preconceptions’
		Gender	**convertirse al*$_m$ *religion*$_f$ ‘convert to the religion’
		Government	**hablando al teléfono* instead of *hablando por teléfono* ‘speaking on the phone’
		Pronoun	**muero de ganas* instead of *me muero de ganas* ‘I’m dying to do sg’
		Incorrect word order	**reputación mala* ‘bad reputation’
	Register errors		#*yo tengo el deseo personal de ser bilingual* ‘I have the personal wish to be bilingual’

Table 13 Classification of errors within the “descriptive” dimension

Finally, as mentioned above, the third main category of the descriptive dimension is constituted by *register errors*. Nevertheless, only one case of register error was identified in the corpus, constituted by the inadequate use of the collocation *I have a wish*, instead of the more suitable expression *me gustaría* ‘I would like to’, in the learner utterance shown in (22), which comes from an essay speaking about the participant’s plans for the future. Note that, although the category of register errors did not appear to be particularly relevant in the case of the learner subcorpus analyzed in this study, it may be more productive with a different dataset.

(22) #*yo* <u>*tengo el deseo personal*</u> *de ser bilingual* lit. 'I have the personal wish to be bilingual'

C. The explanatory dimension

Error categories in the explanatory dimension aim at identifying the source of the error through formulating a hypothesis concerning the production strategy or psycholinguistic process it results from (see Ellis and Barkhuizen 2005, 62). Generally two major types of processes are distinguished in the literature when it comes to explanatory analysis: *interlingual* and *intralingual* errors. The first of these most commonly refer to errors resulting from the influence of the learner's native language, however, in the present study, the possible influence of other languages spoken by the learner was also taken into account. Intralingual errors are assumed to result from learning strategies which can often be observed universally in the production of language learners, independently of their L1. Such strategies involve cases of overgeneralization, simplification, etc., which reflect the learners' imperfect knowledge of the target linguistic system. It should be noted that although aiming to uncover the possible sources of errors is valuable in order to learn about the most common strategies applied by learners, as well as the nature of the language learning process, the results of explanatory error analysis should be treated with caution, given that it is often not possible to identify the source of an error unambiguously. (Ellis and Barkhuizen 2005, 65–66)

In the case of lexical errors, a number of more specific processes or strategies were identified, resulting in two subtypes of errors in the interlingual and three error types in the intralingual error category.

1) The first type of interlingual lexical error is *borrowing*, which involves the use of a lexical item coming from either the learner's L1 or another language spoken by them, different from Spanish. Borrowed items can be used in their original forms, or can be adapted to the phonology (or orthography) of the target language. An example is the use of the Portuguese adjective *inolvidável* instead of Spanish *inolvidable* 'unforgettable' in the combination **experiencia* <u>*inolvidável*</u> 'unforgettable experience' annotated in an essay produced by a learner whose mother is a native speaker of Portuguese.

2) The error type called *extension* refers to the use of an existing L2 lexical item with its meaning and/or combinatory properties being "extended" to take over that of another lexical item. In most cases, one of the meanings of the L1 lexical item corresponds to that of the L2 lexical item used by the learner, leading the learner to “extend” the use this item as an equivalent of the L1 expression in other contexts. An example is the case of the erroneous combination *_gastar tiempo_ used instead of _pasar tiempo_ ‘to spend time’, since the Spanish verb _gastar_ can serve as a translation equivalent of English _spend_ in other contexts, such as e.g. _gastar dinero_ ‘to spend money’, but not when it is combined with temporal expressions. In other cases the L1 lexical item has more than one equivalents in the L2, however the use of these is lexically restricted. An example is the the case of _*gente vieja_ ‘old people’, where the adjective _viejo_ ‘old’, whose use in this context implies certain disrespect, should be substituted by more appropriate _mayor_ ‘elderly’.

A subtype of extension error was established to distinguish cases of extension errors where the incorrect L2 lexical item is _phonetically similar_ to its L1 counterpart, as in e.g. _*doy una marca_ ‘give a mark/grade’ where the Spanish noun _marca_ ‘sign’ instead of _nota_ ‘grade’ is used in analogy with the English noun _mark_, although its meaning is inadequate in the given context. As in the case of the previous example, uses of L1-L2 cognates constitute translation equivalents in certain contexts, while in others, their meanings are unrelated, as with _policía_ ‘police’ in the combination *_policía exterior_ with the intended meaning of _política exterior_ ‘exterior policy’.

3) As it was mentioned above, three intralingual lexical error types were identified. The first of these consists of the use of a non-existent L2 form resulting from _erroneous derivation_ by analogy with other L2 forms. An example is the case of *_enseñanza segundaria_ ‘secondary education’ where the incorrect form *_segundario_ instead of _secundario_ ‘secondary’ is created probably by analogy with _segundo_ ‘second’.

4) _Overgeneralization_ is the second category used to describe intralingual lexical errors. Cases of lexical overgeneralization result from a production strategy whereby the learner substitutes the target form with a lexical item having a more generic semantic content (see also Hussein 1990, 128). An example is the case of the collocate verb in *_empezar una adicción_ lit. ‘start an addition’ instead of _desarrollar_ ‘develop’ with a more specific meaning. In the combination *_malos efectos_, the adjective _malo_ ‘bad’ expressing generic negative evaluation is used instead of _nocivo_ ‘harmful.

5) Finally**,** incorrect uses of L2 lexical items which could not be accounted for by any of the above explanations were tagged as _incorrect lexical selection._ Such was the case of the combination *_saltar un vuelo_ lit. 'jump/skip a flight' used instead of _coger un vuelo_ 'catch a flight'.

EXPLANATORY DIMENSION	Lexical errors	borrowing	*_experiencia inolvidável_ 'unforgettable experience'
		extension	*_gastar tiempo_ instead of _pasar tiempo_ 'spend time'
		extension involving phonetic similarity	*_doy una marca_ instead of _doy una nota_ 'give a mark'
		erroneous derivation	*_enseñanza segundaria_ instead of _enseñanza secundaria_ 'secondary education'
		overgeneralization	*_malos efectos_ instead of _efectos nocivos_ 'harmful effects'
		incorrect lexical selection	*_saltar un vuelo_ lit. 'jump/skip a flight' instead of _coger un vuelo_ 'catch a flight'
	Grammatical errors	interlingual	*_hablando al teléfono_ (from Italian _parlar al teléfono_) instead of _hablando por teléfono_ 'speak on the phone'
		intralingual	*_montó el bus_ instead of _se montó en el bus_ 'he/she got on the bus'

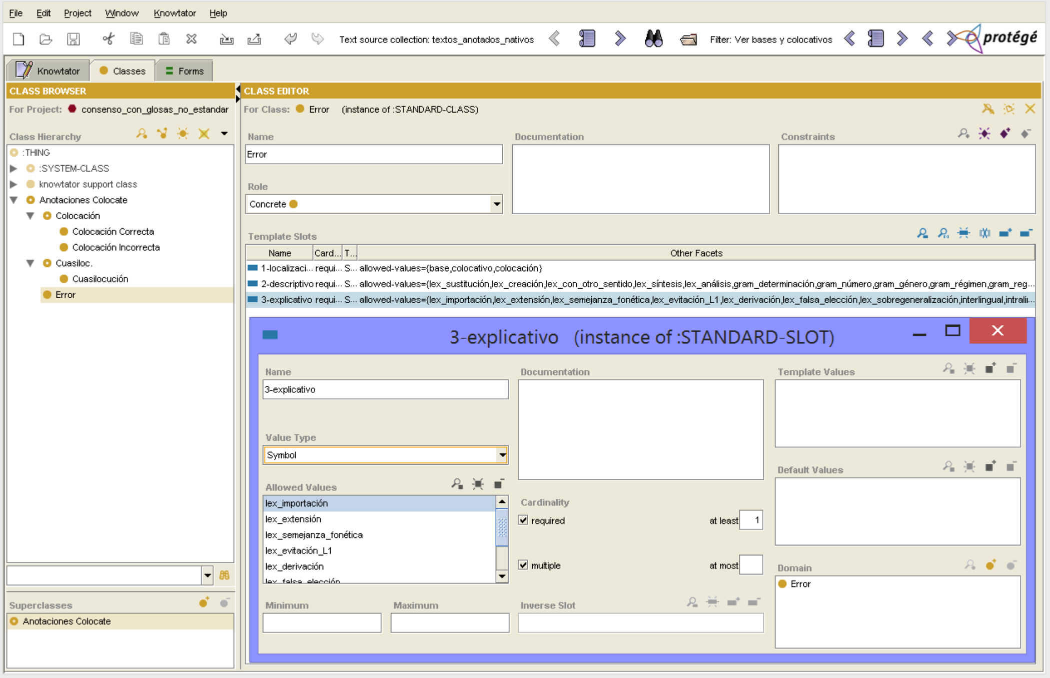
Table 14 Classification of errors within the explanatory dimension

No specific subcategories were foreseen in the explanatory dimension of the error typology for the cases of grammatical and register errors. Hence, these were tagged as either interlingual or intralingual. An example for an interlingual grammatical error, i.e. one that can be attributed to L1 influence is _*hablando al teléfono_ lit. 'speak to the phone' where the erroneous use of the preposition can be derived from _parlare al teléfono_, the equivalent expression in Italian, another foreign language spoken by the learner. The use of the superfluous pronoun in _*se llama la atención_ lit. 'calls the attention' constitutes an example of grammatical intralingual error. Note that, although in the literature production strategies resulting in grammatical errors are described in further detail (see e.g. Ellis and Barkhuizen 2005, 65–66), this was not considered to be necessary for the purposes of the present study.

As it was mentioned above, although identifying the source of learner errors, may render valuable insights, it also involves a considerable amount of speculation. Naturally, the present study is no exception. In the case of a number of erroneous collocations, more than one plausible hypothesis could be formulated regarding the explanation of the error, one generally involving an interlingual, while the other an intralingual error type. These cases were annotated giving preference to the possible influence of the L1. Therefore, for instance, while in the erroneous combination *utilizar una oportunidad* ‘use an opportunity’ could be described as containing an overgeneralization error, since the learner uses a verb with a rather generic meaning instead of more specific *aprovechar* ‘take advantage of’, the collocation was annotated as a case of extension error, given that the collocate used constitutes a translation equivalent of the verb in the homologous L1 combination. As a consequence, all data referring to interlingual or transfer errors obtained from the error analysis, should be interpreted as referring more precisely to *potential* transfer.

4.3.2.2 The corpus annotation tool: Knowtator

The annotation of both the learner and the native subcorpora was carried out through using *Knowtator* (Ogren 2006), a flexible off-the-shelf corpus annotation tool, which is realized as a *Protegé* plugin. The annotation schema, including the error typology described above, was defined in *Knowtator*, as illustrated in Figure 23. The window on the left hand side displays the generic annotation frame, while the two superposed windows on the right show the structure of the specific slots used for error annotation, such as the “location”, “descriptive” and the “explanatory” dimensions, together with the specific error types found under the last category.



Text source collection: textos_anotados_nativos

Filter: Ver bases y colocativos

For Project: consenso_con_glosas_no_estandar

:THING
:SYSTEM-CLASS
knowtator support class
Anotaciones Colocate
Colocación
Colocación Correcta
Colocación Incorrecta
Cuasiloc.
Cuasilocución
Error

Superclasses: Anotaciones Colocate

For Class: Error (instance of :STANDARD-CLASS)

Name: Error

Role: Concrete

Name	Card...	T...	Other Facets
1-localizaci...	requi...	S...	allowed-values={base,colocativo,colocación}
2-descriptivo	requi...	S...	allowed-values={lex_sustitución,lex_creación,lex_con_otro_sentido,lex_síntesis,lex_análisis,gram_determinación,gram_número,gram_género,gram_régimen,gram_reg...
3-explicativo	requi...	S...	allowed-values={lex_importación,lex_extensión,lex_semejanza_fonética,lex_evitación_L1,lex_derivación,lex_falsa_elección,lex_sobregeneralización,interlingual,intrali...

3-explicativo (instance of :STANDARD-SLOT)

Name: 3-explicativo

Value Type: Symbol

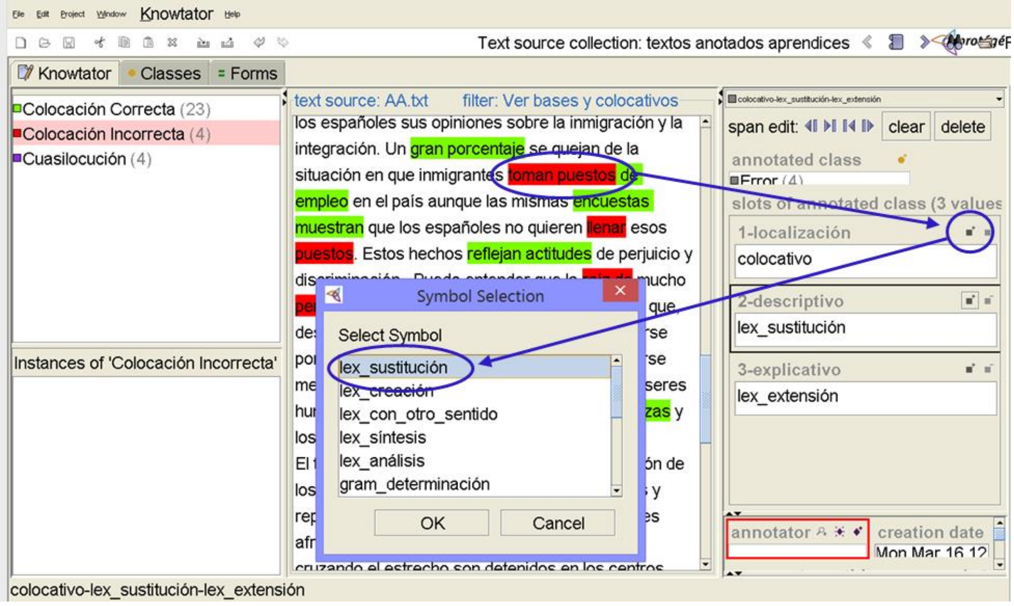
Allowed Values:
lex_importación
lex_extensión
lex_semejanza_fonética
lex_evitación_L1
lex_derivación
lex_falsa_elección

Cardinality:
☑ required at least 1
☑ multiple at most

Domain: Error

Figure 23 Definition of annotation schema in *Knowtator*

Once the annotation schema was established, the corpus had to be loaded in *Knowtator* to start the annotation procedure. Each collocation was annotated through highlighting the corresponding segment of text, selecting the corresponding class (*Colocación correcta* or *Colocación incorrecta*) to create a new entity, and assigning the necessary values in each field to describe the collocation and – when relevant – the error type. Figure 24 shows a screenshot illustrating the selection of the error type when annotating an incorrect collocation. Importantly, the annotation process carried out with *Knowtator* leaves the text files belonging to the corpus intact, with the annotations stored separately. While these are merged and displayed together with the text when using *Knowtator* itself, they can also be exported in XML files to be queried using specific software.


File Edit Project Window Knowtator Help
Text source collection: textos anotados aprendices
Knowtator
Classes
Forms
Colocación Correcta (23)
Colocación Incorrecta (4)
Cuasilocución (4)
Instances of 'Colocación Incorrecta'
text source: AA.txt
filter: Ver bases y colocativos
colocativo-lex_sustitución-lex_extensión
span edit:
clear
delete
annotated class
Error (4)
slots of annotated class (3 values
Symbol Selection
Select Symbol
OK
Cancel
annotator
creation date
Mon Mar 16 12

los españoles sus opiniones sobre la inmigración y la integración. Un gran porcentaje se quejan de la situación en que inmigrantes toman puestos de empleo en el país aunque las mismas encuestas muestran que los españoles no quieren llenar esos puestos. Estos hechos reflejan actitudes de perjuicio y

cruzando el estrecho son detenidos en los centros

lex_sustitución
lex_creación
lex_con_otro_sentido
lex_síntesis
lex_análisis
gram_determinación

1-localización: colocativo
2-descriptivo: lex_sustitución
3-explicativo: lex_extensión

colocativo-lex_sustitución-lex_extensión

Figure 24 Selection of the error type when annotating an incorrect collocation using Knowtator

4.3.3 The annotation procedure

With the aim of studying the collocation production of L2 Spanish learners, the sections of the CEDEL2 corpus described in 4.3.1.2 were manually annotated. The following paragraphs describe the annotation process in detail, as well as some of its main challenges resulting from the notoriously fuzzy interpretation of the notion of collocation.

4.3.3.1 The annotation process

As the first stage of the annotation procedure, each of the one hundred texts in the learner subcorpus was annotated by two native speakers of Spanish independently. The resulting annotations were revised and merged by a consensus annotator, the author of the present thesis. While annotation errors regarding the attributes of collocations were corrected straightforwardly, cases lacking consensus between the two native annotators as to the correctness of a combination the following strategies, similar to those applied by Martelli (2006, 1007–1008) and Nesselhauf (2005, 49–54), were used in order to decide on the correctness of dubious collocations. Firstly, the consensus annotator queried the collocation in the CREA Corpus (Real Academia Española n.d.), and, when at least five occurrences were found, it was considered to constitute a correct combination. Cases that could not be resolved using this method were sent to three independent native annotators, and their correctness was determined according to the majority vote. Finally, dubious annotations and conclusions on merged annotations were discussed in weekly annotation sessions. A similar process was adopted in the case of the annotation of the native subcorpus, although, naturally, error detection played a considerably less prominent role, and consensus annotation focused mainly on the identification of collocations in the corpus.

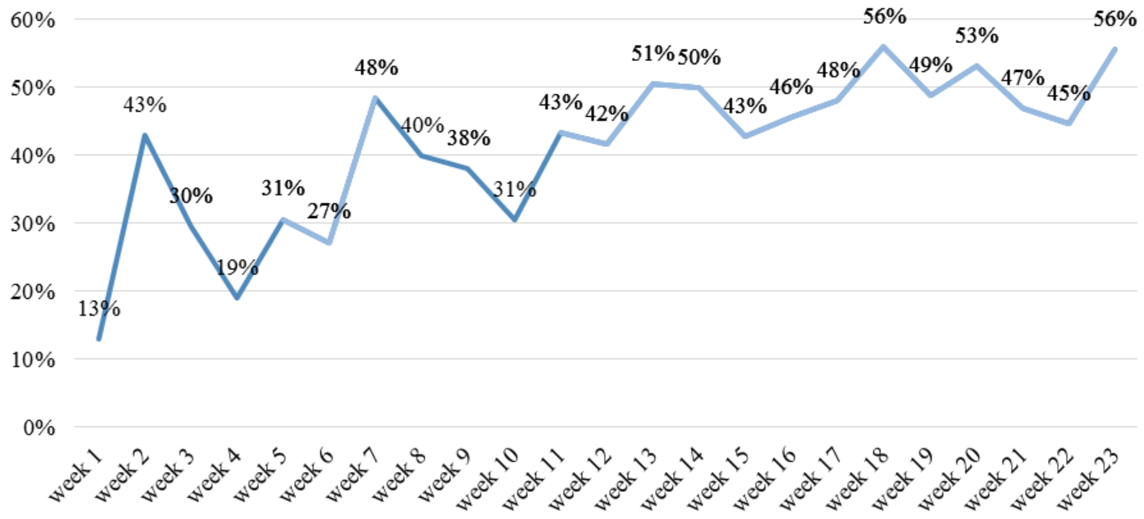


Figure 25 Evolution of inter-annotator agreement throughout the annotation process involving the learner subcorpus

The difficulty of manually annotating collocations can be illustrated by the weekly evolution of consensus between the two native annotators observed during the annotation of the learner corpus. As it can be seen in

Figure 25, despite the well-defined annotation procedure and the weekly sessions dealing with annotation criteria, only a slight increase in inter-annotator agreement was achieved. Inter-annotator agreement in fact remained considerably low throughout the whole of the annotation process: an average of about 30% during the first weeks and average of about 50% over the last weeks. The difficulties observed during the annotation process were found to be generally related to three major issues: the identification of collocations, correction judgment, and error-type annotation. These are discussed in some detail in the following sections.

4.3.3.2 Difficulties posed by identifying collocations

The lack of inter-annotator consensus on identifying collocations in the learner texts could be largely ascribed to the difficulty of establishing clear operational criteria for delimiting the notion of collocation. As a consequence, annotators faced certain difficulty in telling collocations apart from free combinations, on the one hand, and from idioms, on the other.

As it was explained in Chapter 2, the work described in the present thesis adopts the concept of collocation from the ECL framework (see e.g. Mel'čuk et al. 1984; Mel'čuk et al. 1988; Mel'čuk et al. 1992; Mel'čuk et al. 1999), which, in general terms, defines collocations as combinations of two lexical elements where one, the base, determines the selection of the other element, the collocate, to express a given meaning. Meanings of combinations can be represented by LFs with, for instance, the LF Bon codifying the general meaning 'good' (see 2.3.3). The meanings codified by standard LFs are rather generic, and it is assumed that values of a given LF should be included in a lexical entry of a combinatorial dictionary when the semantic characteristics of the base call for the expression of the corresponding meaning (see 2.2.5). For instance, *buena nota* 'good grade' is to be described as a collocation, given that the meaning of the noun *nota* 'grade' calls for a qualification adjective. Nevertheless, in other cases, such as that of the seemingly free combination *buena comida* 'good food', the meaning of the noun does not necessarily require qualification, however, the broadly synonymous combination *comida rica* 'delicious food' involves an adjective with a rather restricted use, since the adjective is prototypically used to refer to good food. According to the criteria adopted in this work, given this restriction, both combinations are to be considered collocations. In other words, when a lexical item imposes a restriction on combining elements expressing a given meaning, the whole range of possible expressions corresponding to that meaning, i.e. possible values of the corresponding LF, were considered as collocates. In the case of *comida* then, the possible values of LF Bon include *bueno* ‘good’, *rico* ‘delicious’ and even *fantástico* ‘fantastic’.

An example for the difficulty of distinguishing collocations from idioms is the case of *darse cuenta* ‘realize’, which constitutes a non-compositional expression given its frozen syntactic structure. It was, nevertheless, mistaken for a collocation by the annotators due to the fact that the verb *dar* ‘give’ is often used in light verb constructions, as in *dar un paseo* ‘take a walk’, *dar consejos* ‘give advice’, etc.

Another issue was that correct collocations frequently passed unnoticed during the annotation process, i.e. they were often perceived as free combinations until an erroneous collocation expressing a similar meaning was encountered. Erroneous combinations naturally make lexical restrictions or mismatches between Spanish and the learners’ L1 more salient. An example for this is the case of *país de origen* ‘country of origin’, which was not annotated as a collocation until the erroneous combination *países maternos* lit. ‘mother(ly) countries’ was found in the corpus. At the same time, annotators tended to perceive any error found in the corpus as a collocation error. For instance, the free combinations **lleno <u>con</u> historia* lit. ‘full with history’ and **recorrimos <u>por</u> la isla* ‘we travelled all over the island’ were both annotated in the first stage of the annotation process by at least one annotator, probably because the preposition errors rendered them more salient[34].

4.3.3.3 Difficulties posed by judging the correctness of collocations

The main issues related to correctness judgment were the individual permissiveness of annotators and the challenges posed by language variation, while the limitations inherent to written text, such as missing intonation pattern, sometimes also caused difficulties in the interpretation of the learner essays.

Native annotators were found to differ in terms of permissiveness towards unusual combinations, which led to a lack of consensus in judging a lexical combination correct or

[34] Recall that combinations consisting of a lexical element and its governed preposition (e.g. *depende de* ‘depend on’), often referred to in the literature as grammatical collocations (see Benson et al. 1986b), are not treated as collocations in the theoretical framework adopted here (see 2.3.1). However, as it is explained in 4.3.2.1, prepositions governed by a member of a collocation (e.g. *tener miedo <u>de</u>* lit. ‘have fear of’) are considered to form part of the expression as a whole, therefore, when erroneous, they are annotated as grammatical collocation errors.

incorrect. At the same time, annotators appeared to be less permissive in accepting creative or unusual language use when dealing with learner essays.

The problem of language variation was noticed especially in the case of collocations typically used in language variants outside of Spain. This issue affected principally the annotation of the learner subcorpus, since, participants contributing to the native subcorpus analyzed were all from Spain, while learners were mainly from the US, and thus many of them had been exposed to variants of Spanish spoken in America. Since native speakers participating in the annotation process were all from Spain, they often judged combinations used in other Spanish-speaking countries to be incorrect, while subsequent verifications using reference corpus data showed that they were in fact in use. These combinations were eventually annotated in the learner corpus as correct collocations, specifying the language variant they belonged to. One example is the case of the combination *hice las reservaciones* 'I made the reservations', which was perceived as incorrect by the annotators, given that in Spain the form *reserva* 'reservation, booking' is used. Differences were also found, for instance, in the use of collocate verbs such as in the combination *tomar clases* lit. 'take classes' (see (23)), where the verb *ir* 'go' is used in Spain in the same context: *ir a clase* lit. 'go to class'.

(23) *Empecé <u>tomando</u> clases de una española* lit. 'I started by taking classes from a Spanish women'

4.3.3.4 Problems posed by error annotation

Three main types of problems were encountered when assigning specific errors to the categories established by the collocation error typology described above. Firstly, given that descriptive error types reflect how an erroneous expression relates to its correct counterpart, in cases when more than one correction was possible, it was sometimes problematic to assign a single error type. For instance, in the sentence shown in (24), the combination *hizo gorditas* can be corrected either for a collocation *ponerse gordas* 'become fat' or a single verb *engordar* 'gain weight'. In the first case the error can be described as the use of an incorrect collocate (*hacer* instead of *ponerse*), while, in the second case, it consists of the use of an erroneous analytical form instead of a single lexical item. In this case, eventually, the first option was included in the definitive dataset, as it was considered as a more natural correction by the native annotators.

(24) *el viaje no *nos hizo gorditas* lit. 'the trip didn't make us fat-diminutive' 'we didn't gain weight during the trip'

Secondly, some incorrect collocation-like combinations identified in the learner corpus were found to be literal translations of L1 combinations that have no collocation equivalent in Spanish. For instance, the erroneous form **humo de segunda mano* corresponds to the English collocation *secondhand smoke*, which can only be translated to Spanish by a complex phrase expressing the same meaning without constituting a phraseological expression: *humo del tabaco de otras personas* ‘smoke from other people’s cigarette’. On the contrary, some expressions used by the learners do not constitute collocations themselves; however the correct form they should be substituted by is a collocation in Spanish. An example for this case can be seen in (25), where the expression containing the copulative verb and the adjective *curioso* ‘curious’ should be reformulated as a collocation: *tengo curiosidad* lit. ‘I have curiosity’. Both of these erroneous expressions were eventually annotated as cases of substitution errors affecting the whole combination.

(25) **estoy curiosa conocerlo* lit. ‘I'm curious to learn about it’

Thirdly, as it was mentioned earlier, in certain cases, the source of the error could not be determined unambiguously. For instance, in the case of the incorrect collocation **hice citas* lit. ‘I made appointments’, it was found feasible to treat the error both as a direct translation of the L1 combination, where the learner seems to have adopted the L2 equivalent of the English support verb, and as a generalization error, whereby the common Spanish support verb *hacer* ‘make/do’ is used instead of the correct and more restricted *concertar* ‘arrange’. In these cases eventually the interlingual error type was preferred.

4.4 Results

In accordance with the above presented research questions, the present section provides quantitative data regarding three main aspects of language learners’ collocation production. Firstly, data concerning the amount of collocations identified in the learner and native subcorpora as well as the amount of different types of collocations is presented. Secondly, the error-rate observed in the case of the learner subcorpus is discussed. Thirdly, data regarding the different collocation error types identified is provided.

As a result of the manual annotation process, a total number of 1825 occurrences of collocations were identified in the CEDEL2 learner subcorpus. These represent a total of 1127 different collocation lemmas. Collocation lemmas were conceived of as consisting of the lemma of the base and the lemma of the collocate, so that e.g. *dimos un paseo* ‘we took a walk' and *daban paseos* 'they took a walk' belong to the same collocation lemma *dar un paseo* 'dar un paseo'.[35] In the native subcorpus, which, as discussed above, was of a smaller extension, 1138 occurrences of collocations, representing 935 different combinations were annotated. The relative frequency of collocations in the learner and the native corpora is 39.31 and 38.02 per thousand words respectively. This does not represent a statistically significant difference as shown by the two proportion z-test, the results of which suggest that the null hypothesis cannot be rejected (z=0.907, p-value=0.36).

However, while the learner and native essays are similar when it comes to collocation density (see also Orol González and Alonso Ramos 2013), native speakers appear to use a larger variety of different combinations. This can be measured by the lemma/token ratio of collocations, which is 0.618, in the case of the learner corpus and 0.822 in the native corpus. Note, however, that while the lemma/token ratio provides information on the amount of different collocations occurring in the two subcorpora as compared to the total amount of collocations annotated, this measure is negatively affected by corpus size. In other words, the lemma/token ratio in a larger corpus tends to be smaller. For a summary of quantitative data regarding corpus size and number of collocations see Table 15.

	LEARNER SUBCORPUS	NATIVE SUBCORPUS
Corpus size (in number of words)	46420	29935
Number of collocation occurrences	1825	1138
Number of collocation lemmas	1127	935
Number of collocations/1000 words	39.31	38.02
Lemma/token ratio	0.618	0.822
Proportion of most frequent 10% collocate lemmas	65.3%	49.7%

Table 15 Summary of data regarding corpus size, number of collocations identified and lexical diversity of collocations

Another way to assess the repertoire of collocations used by learners and native speakers is to examine the occurrences of collocate lemmas as opposed to base lemmas. As it is shown in Table 16, collocates show less lexical diversity than bases in the case of both subcorpora. This can be expected, since, as it was explained in 2.2.5, according to the definition of collocation adopted in this thesis, collocates are selected in function of the

[35] Note that, in order for two collocations to represent forms belonging to the same lemma, they have to be constituted by the same base and collocate lemmas, and correspond to the same syntactic pattern, thus e.g. *pasar tiempo* 'spend time' and *el tiempo pasa* 'time goes by' correspond to two different collocation lemmas.

base to express given meanings, many of which are recurrent. In order to further assess the repertoire of collocates used in each subcorpus, the degree of recurrence of the most frequently occurring collocates was examined. It was found that the 43 most frequent collocates, constituting 10% of collocate lemmas in the learner corpus, were used in 65.3% of all collocations. This proportion is higher than in the case of the native corpus, where the most frequent 46 collocates, constituting 10% of collocate lemmas, were used in 49.7% of collocations, see Table 15. I believe that this piece of data is interesting, since it emphasizes the fact that learners may overuse frequent or default collocates to express given meanings.

	Learner subcorpus		**Native subcorpus**	
	Number of lemmas	**Lemma/token ratio**	**Number of lemmas**	**Lemma/token ratio**
Base	637	0.35	465	0.41
Collocate	433	0.24	567	0.50

Table 16 Lemma/token ratio in the case of bases and collocates in the learner and native subcorpora

In order to obtain quantitative data concerning the types of collocations used, combinations identified in the two subcorpora were categorized according to their syntactic pattern. As it can be seen in Table 17, seven main types of collocations were distinguished: (1) VERB+NOUN$_{COMP}$ (e.g. *ahorrar dinero* ‘save money’), (2) NOUN$_{SUBJ}$+VERB (e.g. *la temperatura se refresca* ‘the temperature cools down’), (3) NOUN+MODIFIER (e.g. *razón principal* ‘main reason’), (4) NOUN+*de*+NOUN (e.g. *paquete de tabaco* ‘pack of cigarettes’), (5) VERB+ADVERB (e.g. *querer sinceramente* ‘love sincerely’) (6) VERB+ADJECTIVE (e.g. *poner nervioso* ‘make nervous’) and (7) MODIFIER+ADJECTIVE combinations (e.g. *sumamente peligroso* ‘extremely dangerous’). An eight category labeled as OTHER was introduced to include collocations not corresponding to any of the previous patterns, such as *lleno de encanto* lit. ‘full of charm’, *entrada en vigor* lit. ‘the coming into effect’ as well as erroneous expressions used instead of a collocation such as *esquiar en agua* lit. ‘ski in water’ instead of *esquí de agua* ‘water ski’ (see the *analysis* error type below). The abbreviation N/A refers to erroneous single word expressions identified in the learner corpus, used instead of collocations, i.e. cases of the error type *synthesis* (see 4.3.2.1).

As the data displayed in Table 17 shows, VERB+NOUN$_{COMP}$ and NOUN+MODIFIER collocations are by far the two most frequently occurring types of combinations in both the learner and the native essays. The relative frequency calculated as the number of occurrences per 1000 words of VERB+NOUN$_{\text{COMP}}$ combinations is 27.122 in the learner subcorpus, as opposed to 20.678 in the native subcorpus. The relative frequency of NOUN+MODIFIER collocations is 8.358 in the learner essays and 11.692 in native texts. The latter type of combinations in most cases consisted of a noun and a modifying adjective, although cases of modifying nouns, prepositional phrases and idiomatic expressions (e.g. *carril bici* ‘bike lane’ *sueño de la vida* ‘life’s dream’, *frío que te mueres* lit. ‘cold that you die’ ‘freezing cold’) were also included in this category.

	Learner subcorpus				**Native subcorpus**			
Pattern	**Number of occurrences**	**Num. of occ.es/10000 words**	**Lemmas**	**Lemma / token ratio**	**Number of occurrences**	**Num. of occ.es/10000 words**	**Lemmas**	**Lemma / token ratio**
V+N$_{\text{COMP}}$	1259	27.122	716	0.569	619	20.678	475	0.767
N$_{\text{SUBJ}}$+V	64	1.379	59	0.922	53	1.771	46	0.868
N+MODIF.	388	8.358	265	0.683	350	11.692	305	0.871
N+*de*+N	50	1.008	38	0.760	63	2.105	58	0.921
V+ADV	41	0.883	28	0.683	23	0.768	22	0.957
V+ADJ	12	0.259	11	0.917	20	0.668	19	0.950
MODIF.+ADJ	6	0.129	5	0.833	3	0.100	3	1.000
OTHER	2	0.043	2	1.000	7	0.234	7	1.000
N/A	3	0.065	3	1.000	-	-	-	-
TOTAL	1825	-	1127	-	1138	-	935	-

Table 17 Distribution of the amount of collocations identified in the learner and native subcorpora according to their syntactic pattern

The remaining types of collocations occur considerably less frequently in both subcorpora. The third and fourth most frequent collocation types are NOUN$_{\text{SUBJ}}$+VERB, with 1.379 occurrences per 1000 words in the learner subcorpus and 1.771 in the native subcorpus, and NOUN+*de*+NOUN combinations with a relative frequency of 1.077 and 2.105 in the learner and the native essays, respectively. The least frequently used types in both subcorpora include VERB+ADVERB combinations, where the verb is modified by either an adverb or an adverbial expression (e.g. *proclamar a los cuatro vientos* ‘announce to all and sundry’), VERB+ADJECTIVE combinations and MODIFIER+ADJECTIVE collocations. Table 17 provides a detailed summary of quantitative data regarding the raw frequency and relative frequency of occurrences of collocations belonging to each group, as well as number of lemmas and type/token ratio.

The chi-square test was applied in order to find out whether the difference in the amount of collocations corresponding to different syntactic patterns used between the learner and the native writers could be explained by chance variation or is likely to constitute a significant difference. The result shows that the overall difference in the distribution of collocations according to syntactic pattern in the two subcorpora can be considered significant, since the null hypothesis can be rejected (χ^2=75.87, p<0.000).[36] As it can be seen in Figure 26 the greatest differences are observed in the case of the two most frequent types of combinations, VERB+NOUN$_{COMP}$ collocations, which are overused by learners as compared to native speakers, and NOUN+MODIFIER collocations, which, on the contrary, are underused in learner essays.

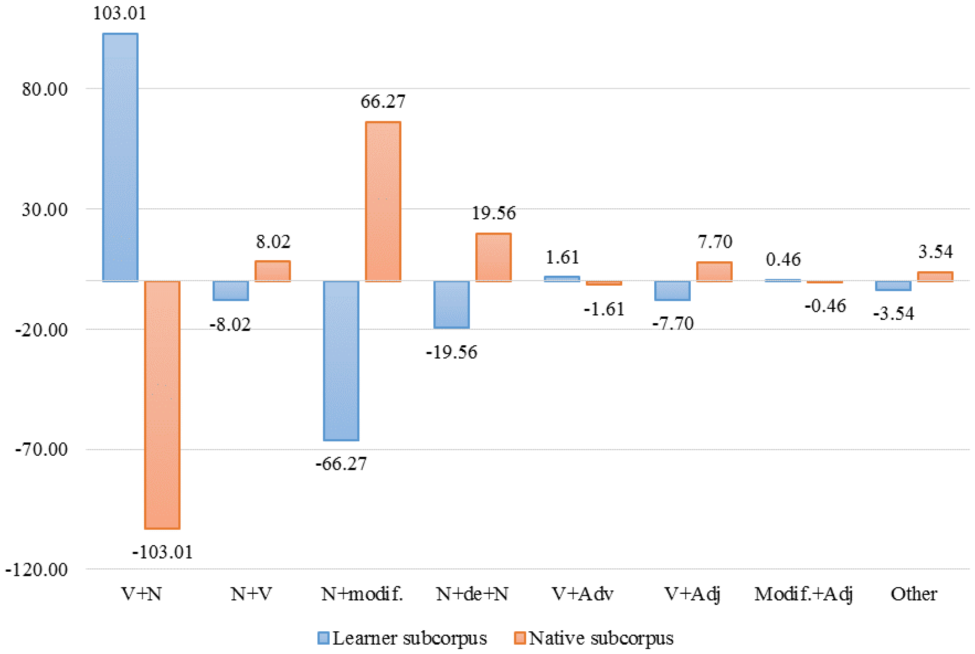


Figure 26 Deviation from the expected number of occurrences per 10000 words in the case of collocations corresponding to the different syntactic patterns studied

Out of the total number of 1825 collocations identified in the learner subcorpus, 1390 occurrences, amounting to 76.16% of all collocations, were judged to be correct, and 435 combinations, constituting 23.84% of all occurrences of collocations, were considered incorrect following the criteria described in 4.3.3.1. Data regarding individual collocation types reveals a similar error rate in the case of the two most frequently occurring syntactic patterns: 22.48% of VERB+NOUN$_{COMP}$ collocations and 23.20% of NOUN+MODIFIER collocations were judged to be incorrect. As it is shown in Table 18, some of the

[36] Since the chi-square test does not yield accurate results with expected values lower than 5, ADJECTIVE+ADVERB collocations and combinations assigned to the category OTHER were grouped together for the purposes of performing the statistical test. Consequently, the contingency table used in the calculation contained seven rows and two columns (df=6). The category N/A was not taken into consideration, since it refers to a specific error type and is not relevant in the case of the native subcorpus.

remaining groups of collocations with a lower number of occurrences present a higher error rate, for instance, 35.94% of NOUN$_{\text{SUBJ}}$+VERB combinations and 40% of NOUN+*de*+NOUN combinations were considered to be erroneous.

Pattern	All occurrences	Correct		Incorrect	
		Num.	%	Num.	%
V+N$_{\text{COMP}}$	1259	976	77.52%	283	22.48%
N$_{\text{SUBJ}}$+V	64	41	64.06%	23	35.94%
N+ADJ	388	298	76.80%	90	23.20%
N+*de*+N	50	30	60.00%	20	40.00%
V+ADV	39	34	82.93%	7	17.95%
V+ADJ	12	7	58.33%	5	41.67%
ADJ+MODIF.	6	3	50.00%	3	50.00%
OTHER	2	1	0.00%	1	100.00%
N/A	3	3	0.00%	3	100.00%
Total	1825	1390	76.16%	435	23.84%

Table 18 Error rate observed in the learner subcorpus according to individual collocation types

In order to give an account of different error types observed in the corpus, instances of errors were considered individually. A total number of 481 error instances were annotated in the 435 erroneous collocations identified in the learner subcorpus. The fact that there is a higher number of error instances than erroneous collocations indicates that, in certain cases, a collocation was tagged for more than one error. For instance, in the case of the erroneous collocation **estamos en vacacion* used instead of *estamos de vacaciones* ‘be on holiday’ a government error and a number error were identified. Two error instances were tagged in a total number of 44 collocations, and three error instances were identified in the case of one collocation in the corpus (see Table 19). In what follows, each of the three dimensions of error analysis represented in the error typology described in 4.3.2.1 will be considered.

Total number of erroneous collocations	435
Collocations with one error	390
Collocations with two error instances	44
Collocations with three error instances	1
Total number of errors instances	481

Table 19 Number of erroneous collocations and error instances in the learner subcorpus

The localization dimension focuses on identifying which element of the collocation is affected by the error. 148 error instances (51.56%) annotated in the corpus affect the *collocate*, the *base* is effected in the case of 170 error instances (35.34%), while 63 collocation instances concern the *collocation as a whole* (see Table 20).

Element affected by the error	Number of error instances	% of all error instances
Base	170	35.34%
Collocate	248	51.56%
Collocation	63	13.10%

Table 20 Distribution of error instances in the localization dimension

As it was explained above, the descriptive dimension of the error typology characterizes errors through making reference either to linguistic categories affected by the error or to ways in which the surface structure of the target expression differs from the erroneous utterance. From among the three main types of descriptive error categories distinguished, lexical errors were the most frequent with 277 error instances, constituting 57.47% of all errors. 203 grammatical errors were identified, which make up 42.12% of all errors, while, as mentioned earlier, one single case of register error was annotated in the corpus (see Table 21).

Descriptive error type	Error instances	
	Number	%
Lexical	277	57.47%
Grammatical	203	42.12%
Register	1	0.21%

Table 21 Distribution of error instances according to the main categories of the descriptive dimension

	Lexical		Gramm.	
	Num.	%	Num.	%
Base	61	35.88%	109	64.12%
Collocate	164	66.13%	84	33.87%
Collocation	52	82.54%	10	15.87%

Table 22 Distribution of lexical and grammatical errors according to the localization dimension

When considering the distribution of lexical and grammatical errors across the categories of the localization dimension, it was found that the base was affected more often by grammatical errors (64.12%) than lexical errors (35.88%), while the collocate was affected more often by lexical errors (66.13%) than by grammatical errors (33.87%). In the case of the 63 errors affecting the collocation as a whole, the majority of errors identified were lexical errors, amounting to 82.54%, while 15.87% of these cases constituted grammatical errors belonging to the word order category. The single register error detected is also considered to affect the collocation as a whole.

Table 23 provides further detail regarding the subtypes of lexical and grammatical error types defined in the descriptive dimension of the error typology. The most frequent type of lexical error is *substitution*, of which a total number of 217 cases were identified in the corpus, making up 78.34% of lexical errors. The second most frequent descriptive lexical error type corresponds to the use of a *correct combination with an incorrect meaning*, which occurred in 11.19% of lexical errors, making it the most frequent type of descriptive lexical error affecting the collocation as a whole. *Analysis* and *synthesis*, the other two lexical error types affecting the whole combination were found to be less frequent, constituting 5.78% and 1.44% of lexical errors, respectively. Finally, 3.25% of lexical errors were annotated as instances of *creation*, i.e. consisting of the use of non-existent L2 lexical elements.

In the case of grammatical errors the most frequent error type affected *government* (16.75%), followed by *gender* (22.17%), *determination* errors (16.75%) and *pronoun* errors (10.34%). The least frequent types of grammatical errors were *word order* (4.93%) and *number* errors (4.43%).

	Descriptive error type	Number of error instances	% of all error instances	% of lexical / grammatical errors
LEXICAL	**Substitution**	217	45.11%	78.34%
	Creation	9	1.87%	3.25%
	Analysis	16	3.33%	5.78%
	Synthesis	4	0.83%	1.44%
	Correct combination with incorrect meaning	31	6.44%	11.19%
GRAMMATICAL	**Determination**	34	7.07%	16.75%
	Number	9	1.87%	4.43%
	Gender	45	9.36%	22.17%
	Government	84	17.46%	41.38%
	Pronoun	21	4.37%	10.34%
	Incorrect word order	10	2.08%	4.93%
	Register	1	0.21%	

Table 23 Distribution of error instances according to the subcategories of the descriptive dimension

The aim of the explanatory dimension of the error typology is to formulate a hypothesis regarding the source of the error. The two main categories established refer to *intralingual* and *interlingual* errors. The number of errors belonging to each of these two categories identified in the corpus is almost identical, with 241 cases of *interlingual* and 240 *intralingual* errors (see Table 24). It is interesting to observe the proportion of errors which likely resulted from L1 influence across grammatical and lexical error types. As shown in Table 25, a greater number of *lexical* errors was considered to be *interlingual* (178) than *intralingual* (99). In contrast, in the case of a larger amount of *grammatical* errors, it was considered that the error resulted from an insufficient knowledge of the L2 linguistic system (141) rather than from the influence of L1 (or another language spoken by the participant) due to the existence of a similar structure (62). In the case of the distribution of *interlingual* and *intralingual* errors across *grammatical* and *lexical* error types, the results of the chi-square test, show that the null hypothesis can be rejected, thus the difference can be considered significant (χ^2=53.275, p<0.000).

Error type	Number of error instances	% of all error instances
Interlingual	241	50.10%
Intralingual	240	49.90%

Table 24 Proportion of interlingual and intralingual errors found in the learner subcorpus

	Interlingual		**Intralingual**	
	Number	%	Number	%
Lexical	178	74.17%	99	41.25%
Gramm.	62	25.83%	141	58.75%

Table 25 Distribution of interlingual and intralingual errors across lexical and grammatical error types

Table 26 shows in more detail the error rate observed in the case of each subtype of *inter*- and *intralingual* error included in the explanatory dimension. As it was explained above, unlike grammatical errors, which were only tagged according to whether they reflected a likely influence of L1 or not, lexical errors were considered in more detail regarding possible production strategies. The most common type of interlingual lexical error was *extension*, with a total of 162 cases amounting to 69.22% of all interlingual errors. Out of these, 54 cases (22.41% of all interlingual errors) were tagged as extension errors where a L2 lexical item was likely chosen due to *phonetic or formal similarity* to its L1 counterpart. *Borrowing* errors were found to occur in a lower number, representing 6.64% of interlingual errors. In the case of lexical intralingual errors, the most frequent error type was found to be *incorrect lexical selection*, with 74 cases constituting 30.83% of all intralingual errors. This is not surprising since the given error tag was used in cases where it was not possible to constitute a hypothesis regarding the production strategy resulting in the error. The remaining error types, *overgeneralization* and *erroneous deviation* errors identified in considerably lower numbers, amounting to 7.92% and 2.5% of intralingual errors, respectively.

			Number of error instances	% of all error instances	% of interlingual / intralingual errors
INTERLINGUAL	LEXICAL	**Borrowing**	16	3.33%	6.64%
		Extension	108	22.45%	44.81%
		Extension due to phonetic similarity	54	11.23%	22.41%
	GRAM.	**Interlingual**	63	13.10%	26.14%
INTERLINGUAL	LEXICAL	**Erroneous derivation**	6	1.25%	2.50%
		Overgeneralization	19	3.95%	7.92%
		Incorrect lexical selection	74	15.38%	30.83%
	GRAM.	**Intralingual**	141	29.31%	58.75%

Table 26 Distribution of error instances according to the subcategories of the explanatory dimension

4.5 Discussion

This section reviews the above presented results in order to provide an answer to the three research questions formulated at the beginning of this chapter.

4.5.1 The amount of collocations used by SFL learners

The first of the research questions referred to examining the extent to which SFL learners use different types of collocations as compared to native speakers. Before considering the data specifically relevant to this question, findings related to the overall amount of collocations observed in both the learner and native subcorpora were presented. As it was shown in the previous section, the raw frequency of collocations found in the learner and native subcorpora differs, however this can be ascribed to the differing size of the two corpora, since the relative frequency of annotated combinations calculated as amount of collocations per 1000 words (39.31 and 38.02 collocations/1000 words in the learner and native subcorpora respectively) was shown to vary across the two subcorpora at only at a negligible rate attributable to chance. In other words, data shows that in the essays constituting the corpus studied here, SFL learners and native speakers seem to have used roughly the same overall amount of collocations.

The lemma/token ratio was used to measure collocation diversity, i.e. the repertoire of different combinations used by native speakers and learners, respectively. A lemma/token ratio of 0.618 was measured in the case of the learner subcorpus, while a somewhat higher ratio of 0.822 was found in the case of the native subcorpus. Although it was noted that these results should be interpreted with caution, they seem to imply that native speakers use a wider variety of combinations. This suggests that learners have a smaller repertoire of lexical combinations than their native peers. Comparable results are offered by Lorenz (1999), who found the type/token ratio to be lower in his study comparing the use of adjective+intensifier expressions in essays produced by EFL learners and native speakers of English.

Returning to the research question concerning the extent to which collocations with different syntactic pattern were used, as it was shown, learners were found to make use of combinations corresponding to each of the seven main patterns, namely, VERB+NOUN$_{\text{COMP}}$, NOUN$_{\text{SUBJ}}$+VERB, NOUN+MODIFIER, NOUN+*de*+NOUN, VERB+ADJECTIVE, VERB+ADVERB and MODIFIER+ ADJECTIVE collocations. The overall distribution of different types of collocations in learner texts was found to differ significantly from that of native speakers. However, as it was suggested, this may be likely attributed to the major differences found in the case of the two most frequent types of collocations, VERB+NOUN$_{\text{COMP}}$ and NOUN+MODIFIER combinations. The data showed an overuse of VERB+NOUN$_{\text{COMP}}$ combinations in the learner corpus, while NOUN+MODIFIER were underused.

When it comes to comparing this data with that obtained in previous studies in the case of EFL learners, as it was noted earlier, Siyanova and Schmitt (2008) found no significant difference in the use of adjective+noun collocations between learners and native speakers, as opposed to the underuse of this type of collocations observed in this study in the case of SFL learners. The case of verb+noun collocations was more extensively studied in EFL learners' production. Here again, data revealing an overuse of this combination type on the part of SFL learners diverges from previous findings in the case of learners of English, since both Howarth (1996) and Laufer and Waldman (2011) reported an underuse of verb+noun collocations in general, while Altenberg and Granger (2001) found EFL learners to use support verb constructions with *make* considerably less than native speakers.

At least part of the overuse of VERB+NOUN$_{\text{COMP}}$ collocations observed in this study can be put down to the use of combinations containing high frequency verbs. See Table 27 for the number of VERB+NOUN$_{COMP}$ collocations containing the ten most frequent verbs in the learner subcorpus. The chi-square test was used to find out whether the different frequencies of occurrence of these collocate verbs can be explained by chance variation.[37] The result shows that, on the contrary, they likely constitute a significant difference, as the null hypothesis can be rejected (χ^2=46.52 > χ^2 critical value=18.48, p<0.000).

	Learner subcorpus		**Native subcorpus**	
	Number of occurrences	**Num. of occ.es/10000 words**	**Number of occurrences**	**Num. of occ.es/10000 words**
tener	343	73.89	97	32.40
hacer	128	27.57	61	20.38
pasar	59	12.71	32	10.69
ir	51	10.99	8	2.67
tomar	46	9.91	33	11.02
dar	41	8.83	34	11.36
hablar	35	7.54	5	1.67
ver	27	5.82	21	7.02
asistir	20	4.31	0	0
comer	20	4.31	0	0
TOTAL	770	-	291	-

Table 27 Verb + noun collocations containing the ten most frequent verbs occurring in the learner subcorpus

The most prominent difference was found in the case of *tener* ‘to have’, which occurs in 343 collocations annotated in the learner corpus as opposed to 97 collocations in the native corpus, some of the most frequent combinations being *tener derecho* ‘have right’, *tener problema* ‘have a problem’, and *tener oportunidad* ‘have an opportunity’. Combinations with *tener* were also frequently used – on occasions erroneously – to describe the caracteristics of an object, a person, a place, etc., as in **tiene una historia muy rica* lit. ‘it has a rich history’, i.e. ‘it’s a historical place’, **tiene mucha lluvia* lit. ‘it has a lot of rain’, i.e. ‘it is a rainy place’, etc. It can be assumed that the use of VERB+NOUN$_{COMP}$ combinations containing frequent verbs constitutes a strategy to circumvent more sophisticated vocabulary items or complex grammatical structures. These observations appear to be in line with those of other researchers claiming that language learners tend to

[37] Since no occurrences of *asistir* and *comer* as collocate verbs were recorded in the native corpus, the statistical test was carried out using the frequencies of the eight most frequent collocate verbs listed in Table 27. Consequently the contingency table used to calculate the χ^2 value contained eight rows and two columns (df=7).

show a preference towards using high frequency combinations and/or combinations consisting of high frequency lexical items (see Durrant 2008; Kaszubski 2000; Nesselhauf 2005). Note that another possible explanation for the overuse of patterns was provided by Lorenz (1999), who, as mentioned in the previous chapter, explains the overuse of adjective intensification in argumentative essays written by L1 German EFL students through the particular student group's tendency to information over charge. The claim made by this author suggests that over and underuse phenomena are not necessarily explained by learners' lower competence, but should also be considered from the point of view of the frequency of use of given structures in the learners' L1.

4.5.2 The amount of erroneous collocations identified in the corpus

The second research question was concerned with the overall error rate observed in the case of the collocations annotated in the learner subcorpus, as well as the error rate established in the case of combinations corresponding to different syntactic patterns. Results discussed above show that 23.84% of collocations identified in the learner corpus was judged to be erroneous. The proportion of erroneous collocations established in the present study is considerably lower than the over 50% error rate obtained by Pérez Serrano (2014) using essays produced by SFL learners at beginner and elementary levels. This seems to show that language learners' collocational competence improves together with their language proficiency, at least as regards the amount of correct combinations produced. However, in order to obtain more accurate data, naturally, an analysis of essays produced by learners corresponding to different proficiency levels following strictly the same criteria should be carried out.

As for individual collocation types, as shown in Figure 27, a similar error rate was found in the case of the two most frequently occurring categories, VERB+NOUN$_{\text{COMP}}$ and NOUN+MODIFIER combinations, with 22.48% and 23.20% erroneous collocations respectively. Considerably higher error rates of 40% and higher were found in the case of NOUN+*de*+NOUN, VERB+ADJECTIVE and MODIFIER+ADJECTIVE combinations, while the lowest error rate was detected in the case of VERB+ADVERB collocations, with less than 18% of instances judged to be incorrect. Although error rate can be indicative of the difficulty of each type of combination, constituting information of pedagogical value, given the low number of overall occurrences in the case of the latter patterns, it is hardly possible to draw definitive conclusions from the data. In a similar attempt to establish a difficulty scale using collocation patterns, in this case, on the basis of a collocational competence test administered to EFL learners, Moreno Jaén (2009) found subjects to perform slightly better on test items involving adjective+noun combinations as compared to noun+verb collocations, although results of the statistical test applied indicated that the difference could be the result of chance distribution.

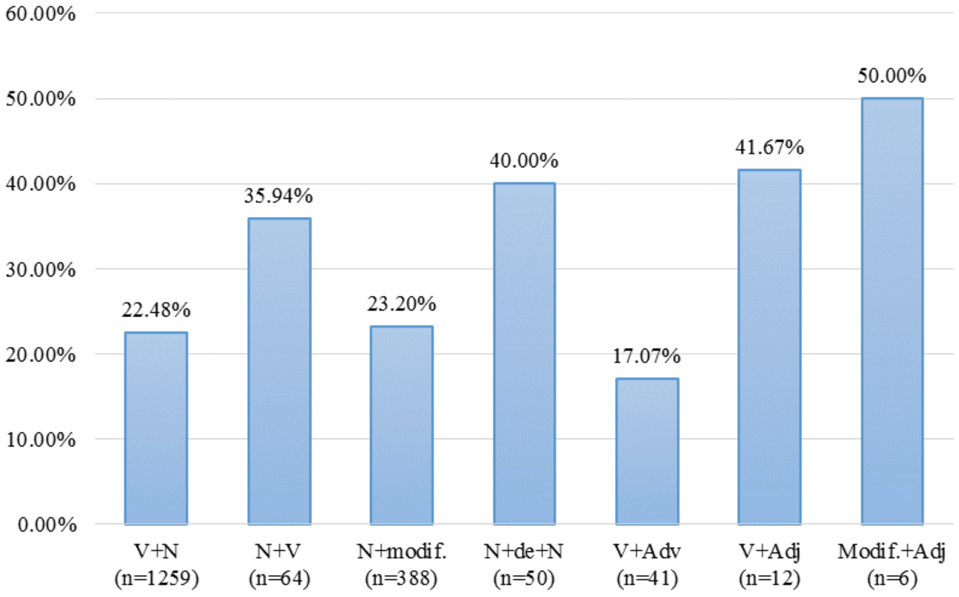


Figure 27 Error rate observed in the learner corpus across collocations corresponding to different syntactic patterns

4.5.3 Error types identified in the corpus and pedagogical implications

Errors identified in the present corpus study were described according to three main sets of criteria. The first of these concerned the element of the collocation affected by the error, the second described the nature of the error, making use of categories representing a combination of linguistic and surface structure taxonomies, finally a third set of criteria was used to explore the source of the error. In accordance with the third research question, this section summarizes the outcomes of the analysis of errors focusing on their pedagogical implications.

The element of the collocation most frequently affected by errors was found to be the collocate. Importantly, this observation also holds if we focus only on lexical errors, in fact, while most errors affecting the collocate were lexical errors, in the case of the base, grammatical errors were more common, see Figure 28.

This seems to be in accordance with the notion of collocation adopted in this thesis, establishing that, while the base constitutes an autonomous element, it conditions the choice of the collocate (see 2.2.4.4 and 2.2.5) Consequently, it has been hypothesized that the production on collocations poses a problem for the language learner mainly when it comes to selecting the collocate (cf. Hausmann 1979; Hausmann 1989). Following this idea, lexical errors concerning the base, such as in e.g. *afecto malo* lit. 'bad affection' instead of *efecto* 'effect', can be considered as no different from lexical errors affecting any single lexical element, while genuine lexical errors of the collocation would be exclusively those concerning the erroneous choice of the collocate, assumed to pose an additional difficulty, given the phenomenon of restricted lexical selection. Some examples for lexical errors affecting the collocate identified in the CEDEL2 corpus are *<u>*empezar*</u> *una familia* 'to start a family' (instead of *formar una familia* lit. 'to form a family'), **entrar la cola* lit. 'enter the queue' (instead of *ponerse a la cola* 'join a queue'), **interrumpir una ley* lit. 'to interrupt a law' (instead of *violar una ley* 'to break a law') and **tener picnic* 'to have picnic' (instead of *hacer picnic* lit. 'to do picnic').

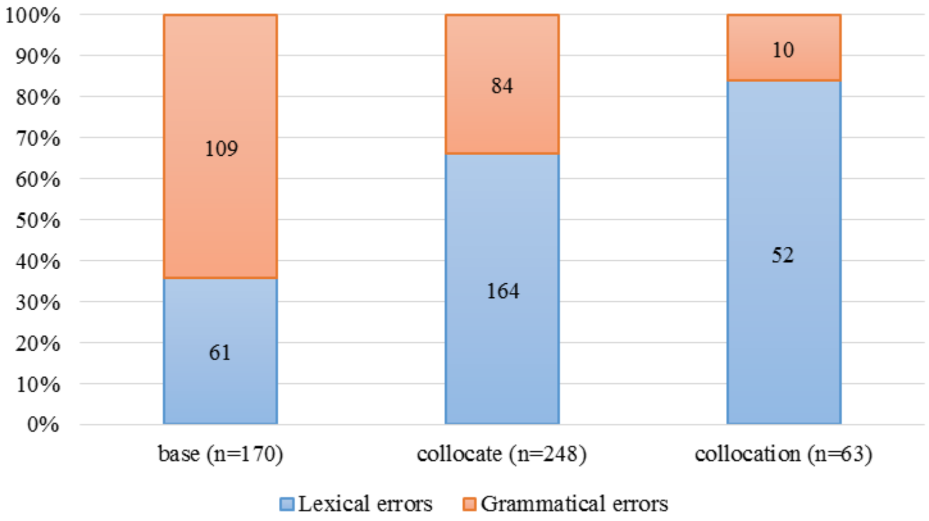


Figure 28 Proportion of lexical and grammatical errors affecting the elements of the collocation

As it can be observed in the above examples, two main types of lexical collocate errors can be identified. In the case of *<u>*empezar*</u> *una familia* 'to start a family' and *tener picnic* 'to have picnic' the combinations produced by learners are congruent with existing L1 combinations, while in **interrumpir una ley* lit. 'to interrupt a law' and **entrar la cola* lit. 'enter the queue' the collocates chosen by the learners do not seem to be congruent with L1 expressions, but have a similar or related meaning to that of the intended L2 item.

These two types of errors, corresponding to the *substitution* and *erroneous lexical choice* categories of the error typology respectively, seem to reveal two main production strategies applied by language learners. Firstly, when the erroneous collocate produced is likely to be the result of L1 transfer, it can be hypothesized that the strategy followed by the learner was to pick an L2 item that constitutes a translation equivalent of the L1 collocate used in the corresponding collocation. Secondly, when it does not seem feasible that the combination produced would result from transfer, it can by hypothesized that the strategy followed implies the choice of an L2 item which is judged suitable by the learner to convey the meaning intended by the collocate.

Note that what is often not taken into consideration in the literature is that the existence of these two possible production strategies and/or the dominance of either does not allow to draw strong conclusions regarding the underlying motivation when relying solely on corpus data. It is generally assumed that lexical collocation errors, i.e. instances of erroneous choices of a collocate, result from language learners' lack of awareness concerning the nature of collocations (see e.g. Farghal and Obiedat 1995). As a consequence, both the base and the collocate are chosen as if they were entirely autonomous, or using Hausmann's terminology, "autosemantic" elements. This is in line with conclusions made mainly on the basis of the results of psycholinguistic experiments by e.g. Wray (2002), who claims that foreign language learners tend to focus on individual words, failing to notice and acquire formulaic chunks, as well as assumptions that language learners' production is dominated by the open choice principle as opposed to the idiom principle (see e.g. Granger 1998; Durrant 2008), see 3.2.2. Nevertheless, it is also possible to hypothesize that – at least in some cases – learners can rely on L1 transfer to fill lexical gaps, i.e. to supply collocates they are not familiar with in the L2, despite the that they are to some extent aware of the fact that the resulting expressions are not necessarily native-like.

We have already seen that learners' heavy reliance on their native language when producing collocations is one of the most recurring observations in the literature. Nesselhauf (2005, 181) claims that L1 influence is the likely source of error in over 50% of erroneous combinations produced by learners in her corpus. Nevertheless, L1 transfer does not necessarily result in an error. Accordingly, learners have also been found to overuse native-like L2 collocations which are congruent with their L1 (Granger 1998, 150–151; Kaszubski 2000, 243). Furthermore, it has been observed that learners' reliance on their native language depends on the perceived distance between the L1 and the L2 (Kaszubski 2000, 246; Uriel Domínguez 2014, 49).

In order to assess the relevance of different strategies involving L1 influence in the case of lexical errors, it is necessary to consider the rate of likely L1 transfer in the dataset. As it was shown in the previous section, while the overall proportion of *interlingual*, i.e. transfer errors, and *intralingual* errors was 50%-50%, the proportion of *interlingual* errors was higher than that of *intralingual* errors in the case of *lexical* errors, with a total of 178 (64.26%) lexical error instances likely resulting from L1 transfer. Note that while this data shows reliance on L1 forms as the more frequently used production strategy, it should be considered with caution, given the criterion followed in the annotation process to tag potentially ambiguous errors as a likely result of transfer whenever that hypothesis seemed feasible, which might have resulted in certain bias towards transfer errors.

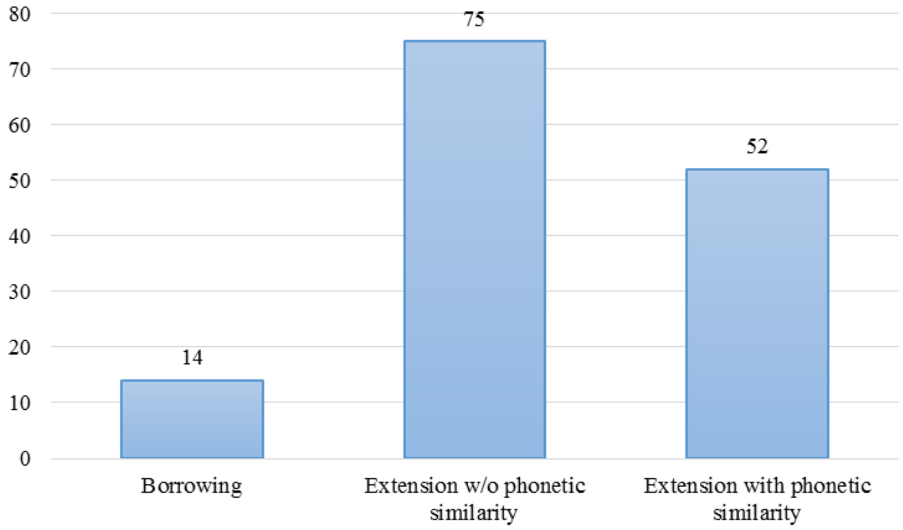


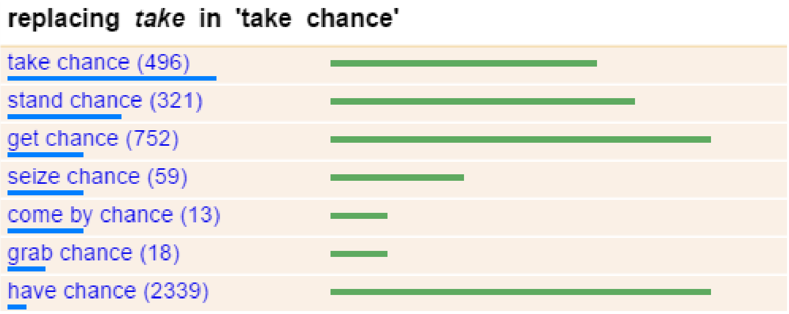
Figure 29 Subtypes of interlingual lexical errors affecting the base and the collocate (n=141)

Among the lexical transfer errors affecting an element of the collocation (the base or the collocate), but not the combination as a whole, the more commonly observed error types involved the use of L2 lexical items constituting a translation equivalent, or assumed translation equivalent of an L1 lexical item, while the incorporation of an L1 lexical item, was limited to the case of 14 error instances, see Figure 29. Half of the latter cases involved the use of Portuguese words instead of Spanish equivalents, and were produced by a participant one of whose home languages was Portuguese. In some other cases L1 forms were adapted to Spanish morphology, as in **misinterpretaciones* instead of *malas interpretaciones* ‘misinterpretations’. As for uses of an L2 translation equivalent of an L1 lexical item, the L2 item was phonetically similar to the L1 form it was assumed to originate from in the case of 52 error instances (e.g. *extiende la vida* ‘it extends life’ instead of *alargar la vida* lit. ‘lengthen life’), while in the remaining 75 cases the L2 element did not bear a formal similarity to its L1 counterpart.

So far we have focused exclusively on the nature of lexical errors identified in the corpus. The descriptive dimension of the error typology used in the present study, however, also contained a number of categories serving to classify grammatical errors relevant in the correct formulation of the collocation. As it was shown in the previous section, over 40% of all error instances annotated in the corpus were grammatical errors.

The pedagogical implications of this observation are twofold. On the one hand, the high number of grammatical errors highlights the importance of providing a sufficiently detailed and, at the same time, transparent description of the characteristics of collocations in reference materials, such as collocation dictionaries. While, as we saw in 3.4.2.1, these dictionaries are often confined to providing lists of co-occurring lexical items, it seems to be important to specify the use of articles, cases of obligatory use of plural or singular noun forms as well as government pattern or obligatory word order. On the other hand, instructors should emphasize these characteristics when teaching collocations to promote precision in language production, while learners’ skills in obtaining grammatical information from dictionaries as well as deducing usage patterns from example sentences or other type of L2 input should also be enhanced.

Another important point, when it comes to the high rate of grammatical collocation errors identified in the corpus, is related to collocation checker tools discussed in 3.4.2.3.C. The fact that these tools focus primarily on the lexical elements of the collocation has been already mentioned. Namely, *Collocation Checker* (Chang et al. 2008) only allows the user to introduce a combination in the form of a bare infinitive followed by a non-inflected noun, often constituting incorrect formulations (e.g. *lay table, take chance*), while correction suggestions in the same tool, similarly to the list of combinations offered in *Just the Word* (Edmonds n.d.) appear in the same format (see Figure 30). This way of displaying collocations may be misleading to users, who are, in addition, left to their own devices when it comes to exploring any additional information required for the correct use of the combination, e.g. through the corpus examples provided.

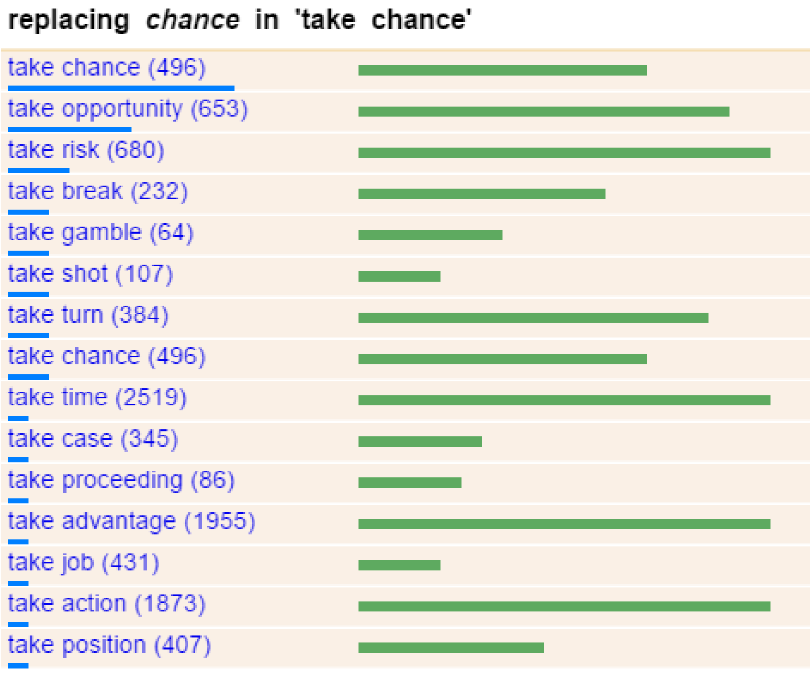


replacing *take* in 'take chance'

take chance (496)
stand chance (321)
get chance (752)
seize chance (59)
come by chance (13)
grab chance (18)
have chance (2339)

Figure 30 Alternative verbs to be combined with the noun *chance* offered by *Just the Word*

When it comes to lexical errors, *Collocation Checker* only anticipates cases of erroneous collocates in verb+noun combinations, consequently correction suggestions are limited to alternative verbs that a given noun can be used with. As we have seen in the corpus data, learners often also choose the lexical item constituting the base erroneously, in which case this tool would be of no help. On the contrary, as it was already discussed, *Just the Word* does offer alternative options for the base, while maintaining the collocate (see Figure 31). In the case of the erroneous collocation **lograr un gol* lit. 'achieve a goal (in sport)' instead of *lograr un objetivo* 'achieve an aim', for instance, an analogous system working with Spanish would likely provide the desired expression. Naturally, this solution is not necessarily satisfactory, such as e.g. in the case of **salir de su posición* lit. 'leave of his position' instead of *dejar su trabajo* 'quit his job', where both the base and the collocate are erroneous.



replacing *chance* in 'take chance'

take chance (496)
take opportunity (653)
take risk (680)
take break (232)
take gamble (64)
take shot (107)
take turn (384)
take chance (496)
take time (2519)
take case (345)
take proceeding (86)
take advantage (1955)
take job (431)
take action (1873)
take position (407)

Figure 31 Alternative nouns to be combined with the verb *take* offered by *Just the Word*

Further challenging cases for collocation checkers would be constituted by the different error types defined in the error typology as affecting the collocation as a whole. An example for the use of an otherwise *correct combination with an incorrect meaning* is shown in (26), where, in order to express the intended meaning, the combination *encender el fuego* 'light the fire' should be substituted for *prender fuego* 'set fire to sg'. Other similarly problematic cases are *analysis* errors involving the use of an incorrect collocation-like expression to convey a meaning that should be correctly expressed by a single word (see (27)), the use of collocation-like erroneous expressions that should be substituted by a non-idiomatic utterance (28), and *synthesis* errors involving the use of incorrect single-word forms standing instead of a collocation (29). Finally, the occurrence of L2 non-words, as we have seen in the case of *borrowing* as well as *creation* errors, when a new L2 item is coined usually through a derivation process, may also cause complications in the case of tools relying solely data coming from an L2 reference corpus (see (30)).

(26) y ella *encendio el fuego que los quemaron (instead of *prender fuego* 'set fire to sg') lit. 'and she lit the fire that burnt them [the mother and her lover]'

(27) *haciendo de cotilleo lit. 'making gossip' (instead of *cotillear* 'to gossip')

(28) *humo de segunda mano 'secondhand smoke' (instead of *humo del tabaco de otras personas* 'smoke from other people's cigarette')

(29) El día siguente, nosotros nos despedimos y *gracias […] (instead of *dar las gracias* 'to thank') lit. 'the following day we said goodbye and thanks'

(30) Hay tres tareas en el concurso que Harry necesita *completir (instead of *llevar a cabo* 'to complete') lit. 'There are three tasks in the contest that Harry has to complete'

4.6 Summary

This chapter described a learner corpus study whose aim was to explore SFL learners' collocation production, which was deemed necessary both in order to contribute to filling the gap given the lack of such studies on Spanish learner language, as well as to provide empirical basis to proposing design criteria for collocation learning tools.

The study made use of a portion of the CEDEL2 corpus comprised of 100 learner essays and 103 texts produced by native speakers, in which collocations had been manually annotated. The annotation process was described in detail, with special emphasis on the error typology used to classify erroneous collocations. The typology aimed at capturing three different aspects of collocation errors constituted by the location of the error, descriptive and explanatory analysis. The main difficulties encountered during the manual collocation annotation process, related to the identification of collocations, correction judgment and error annotation, were also discussed.

One approach to analyzing learners' language production is comparing their output to that of native peers. In this respect, the results of the present corpus study showed that the essays produced by SFL learners did not differ from those written by native speakers when it comes to the overall amount of collocations, measured through the relative frequency per thousand words. However, as it can be expected, learners were found to use a smaller repertoire of combinations than native speakers of Spanish, as indicated by the lower lemma/token ratio.

When it comes to the types of collocations used, it was observed that the distribution of the seven main types of syntactic patterns studied differed significantly across the learner and the native subcorpora. It was noted that this can likely be attributed to the difference in the use of the two most frequently occurring types of combinations, VERB+NOUN$_{\text{COMP}}$ and NOUN+MODIFIER collocations. While the first of these was found to be overused by language learners – in part due to the high amount of combinations containing high frequency verbs, especially *tener* 'to have' –, NOUN+MODIFIER collocations were found to be underused in the learner corpus, as comparted to native writing.

Learner essays were considered in more detail with respect to erroneous collocations. Of all collocations identified in the texts produced by learners, 23.84% were judged to be erroneous. At the same time, it was noted that the data corresponding to individual collocation types obtained from the corpus does not allow for definitive conclusions as to their comparative difficulty as defined by error rate.

A more detailed error analysis revealed that most lexical errors identified in the corpus affected the collocate, i.e. resulted from learners' insufficient knowledge concerning lexical combinatory restrictions, as it could be expected. Concerning the source of errors, results are in line with those of previous studies in showing that most lexical errors likely resulted from L1 transfer. It was also noted, however, that, although this latter observation has certain importance regarding the production strategies applied by learners when it comes to collocations, it is not conclusive when it comes to considering whether or the extent to which language learners are aware of the phenomenon of lexical combinatory restrictions. It was suggested that the fact that learners make use of L1 transfer when producing collocations does not necessarily mean that they are not aware of the restricted nature of a combination, they simply might not have other means of expressing the desired meaning. Finally, it was also argued that the fine-grained typology used for error annotation not only allowed to describe collocation errors identified in the learner corpus in detail, but it also served to foreground aspects of collocation production that may not have received sufficient attention, such as the cases of grammatical errors, errors concerning the base, whole collocation errors or target language non-words.

Chapter 5. Testing collocation learning resources

5.1 Introduction

The development of effective language learning resources not only requires sufficient theoretical foundation and insight into language learners' knowledge as well as problem areas concerning the target linguistic phenomena. It is also necessary to be aware of how learners constituting the target user group can potentially interact with a given type of tool to enhance their production, comprehension and further their knowledge. However, as it was noted in 3.4.2, while SFL learners have considerably less resources at their disposal than learners of English, there is also a notable lack of empirical studies focusing on language learners' interactions with different tools that can be exploited for collocation learning, including collocation dictionaries, language corpora and online learning tools. Hence, the present chapter examines the use of two types of resources with the aim of assessing how successfully SFL learners can manipulate them, and formulating design recommendations for the improvement of existing resources, as well as for the development of future learning tools. Accordingly, the first part of the chapter describes a usability test involving *Diccionario de colocaciones del español*, a freely available online collocation dictionary of Spanish (see 3.4.2.1), while the second part presents an experimental study whose aim is to assess SFL learners' ability to correct collocation errors with the help of concordance feedback.

5.2 Testing an electronic collocation dictionary interface

The following sections describe a usability study that tests the online interface of the *Diccionario de colocaciones del español* (DiCE, Alonso Ramos 2004). This dictionary was conceived with the purpose of providing a theoretically well-founded and detailed description of Spanish collocations. The usability test described here was carried out in order to assess to what extent this dictionary constitutes a useful tool for different target user groups, and to identify design issues that need to be dealt with in future versions of the interface.

5.2.1 Aims of the study

Dictionary use can be studied from different perspectives. While many studies aim at finding out what purposes dictionaries are used for, what knowledge or abilities dictionary users have and need, or how dictionaries contribute to language learning, as we have seen in 3.4.2.1.E, Heid (2011) proposes to adapt *usability testing*, as understood in information science, to electronic dictionaries. This line of research implies testing dictionaries at the level of functionality, much like in the case of other kinds of software tools. Studies that apply usability testing methodology include Heid and Zimmerman (2012), comparing different types of access routes in mock-up dictionary interfaces, and Hamel (2012), providing a detailed description of a usability experiment with a dictionary prototype concentrating on lexical selection, combination and paraphrase. In addition, as it was noted earlier, a number of studies pursued similar goals in the case of paper and electronic collocation dictionaries. Komuro (2009) and Lew and Radłowska (2010) carried out experimental studies to investigate the use of the *Oxford Collocations Dictionary 1st edition* (Crowther et al., 2002), while Alonso Ramos (2008) compared learners' success in interpreting dictionary entries retrieved from two Spanish combinatory dictionaries *Práctico* (Bosque 2006) and the DiCE itself (see 3.4.2.1.E).

The aims of the usability study designed to test the DiCE interface can be summarized in the following five research questions:

1) How effective and efficient are user interactions with the dictionary interface in general?
2) How effective and efficient are user interactions when searching for different types of information?
3) Do participants corresponding to different user profiles differ in terms of efficiency and effectiveness of interactions?
4) What usability problems can be identified within the DiCE interface?
5) To what extent are users satisfied with DiCE?

5.2.2 The dictionary interface tested: Description of DiCE

DiCE was designed in accordance with the postulates of the Explanatory Combinatorial Lexicography (ECL, Mel'čuk et al., 1995; see 2.3.3), consequently, as it was discussed in 3.4.2.1.B, the design of lexical entries is oriented towards language production, with collocates listed in the entries of corresponding bases.

In order to offer more dynamic access to the information stored in the DiCE database, in addition to *Acceso al diccionario* (Dictionary access), where lexical entries can be browsed in a manner similar to a paper dictionary, the current user interface also incorporates various advanced search options. Each of these was conceived to provide the user with a more direct path of access to a specific type of information. Since one of the main objectives of the usability test was to assess the functionality of the different search options, the following paragraphs provide a brief description of these. Figure 32 shows an illustration of the DiCE user interface.

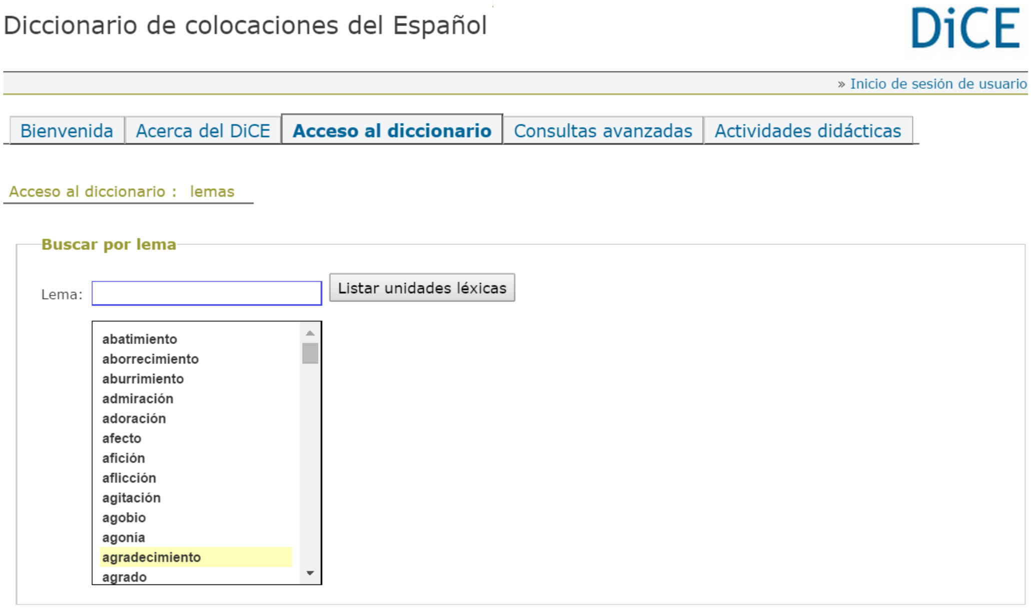
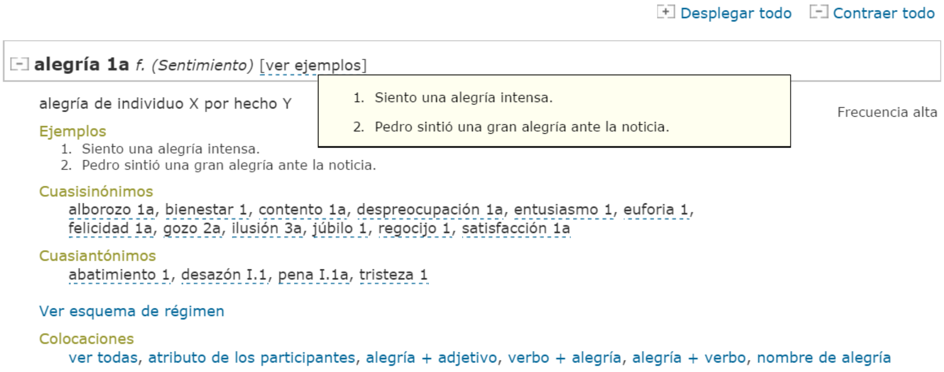


Figure 32 Illustration of the DiCE user interface with the *Acceso al diccionario* option

The DiCE interface contains two main modules through which users can access the dictionary database: *Acceso al diccionario* and *Consultas avanzadas* (Advanced search options). *Acceso al diccionario* (see Figure 32), as mentioned above, provides a traditional collocation dictionary type access to the combinatorial information stored in DiCE. The entry of each lemma contains the subentries of its corresponding lexical units (LU), while the collocations of each LU are grouped according to their syntactic pattern and semantic content. *Consultas avanzadas* consists of four independent search options: 1) *¿Qué significa?* 2) *Ayuda a la redacción*, 3) *Consultas directas* and 4) *Consultas inversas*.



Desplegar todo Contraer todo

alegría 1a *f. (Sentimiento)* [ver ejemplos]

1. Siento una alegría intensa.
2. Pedro sintió una gran alegría ante la noticia.

alegría de individuo X por hecho Y

Frecuencia alta

Ejemplos
1. Siento una alegría intensa.
2. Pedro sintió una gran alegría ante la noticia.

Cuasisinónimos
alborozo 1a, bienestar 1, contento 1a, despreocupación 1a, entusiasmo 1, euforia 1, felicidad 1a, gozo 2a, ilusión 3a, júbilo 1, regocijo 1, satisfacción 1a

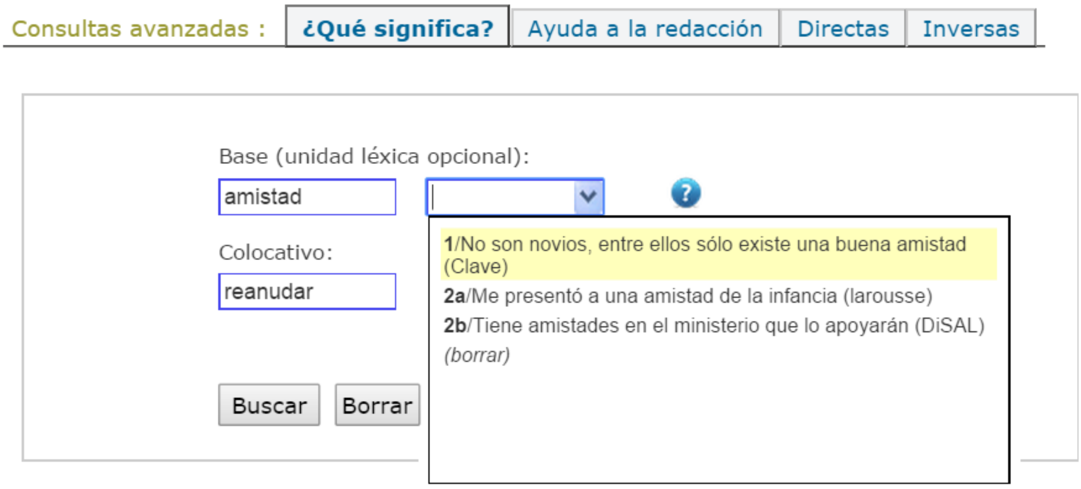
Cuasiantónimos
abatimiento 1, desazón I.1, pena I.1a, tristeza 1

Ver esquema de régimen

Colocaciones
ver todas, atributo de los participantes, alegría + adjetivo, verbo + alegría, alegría + verbo, nombre de alegría

Figure 33 Lexical entry for a lexical unit in *DiCE*

The first advanced search option, *¿Qué significa?* (What does it mean?), is oriented to reception. The user is prompted to introduce a base (e.g. *amistad*) and a collocate (e.g. *reanudar*), and the search results show the entry of the corresponding collocation, where the meaning of the combination is specified by a gloss, and usage examples are also shown (see Figure 38). Note that, the user can further specify the query by indicating the LU of the base. See Figure 34 for an illustration of this search option.



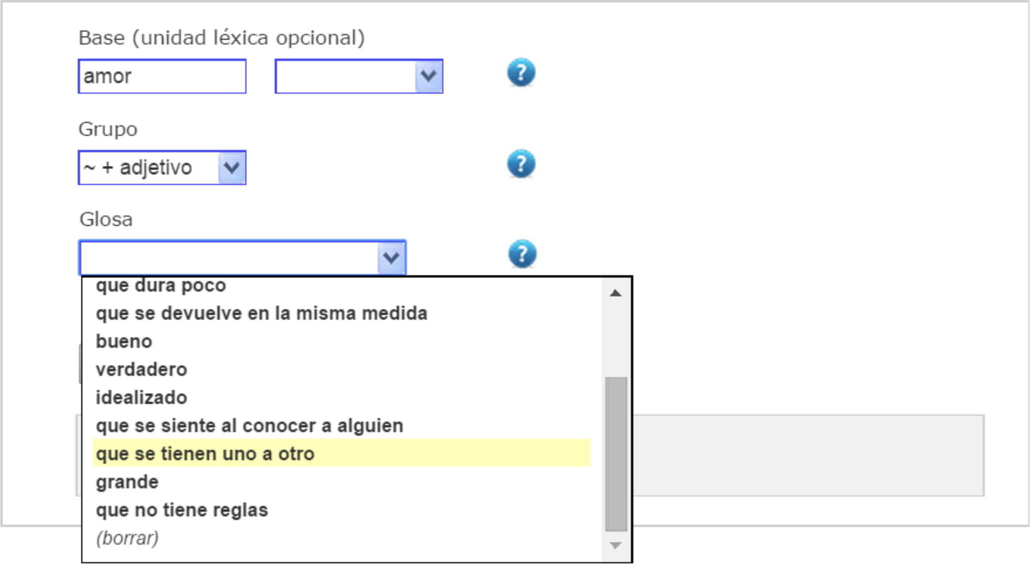
Base (unidad léxica opcional): amistad

Colocativo: reanudar

1/No son novios, entre ellos sólo existe una buena amistad (Clave)
2a/Me presentó a una amistad de la infancia (larousse)
2b/Tiene amistades en el ministerio que lo apoyarán (DiSAL)
(borrar)

Figure 34 The *¿Qué significa?* search option

The *Ayuda a la redacción* (Writing aid) search option is oriented to language production. It allows the user to find collocates of a given base, corresponding to a specific part of speech, and a meaning represented by a semantic gloss. For instance, one can search for an adjective to be combined with the noun *amor* 'love' to express the meaning 'felt for one another'. As in the case of the above search option, the user has the option of specifying the LU of the base. Once the query is complete, the entries of corresponding collocations are displayed. In the case of the specific query given here as an example and illustrated in Figure 35, the collocation returned by the interface is *amor mutuo* ‘mutual love’.



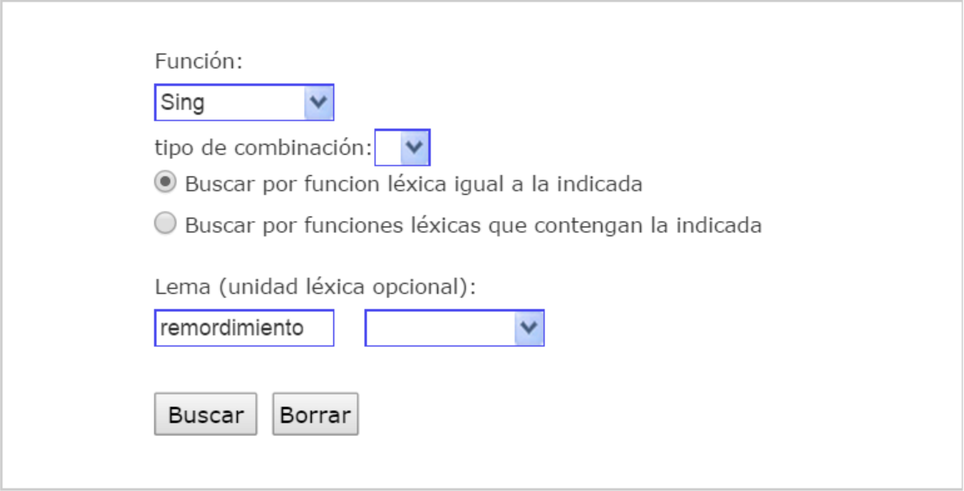
Base (unidad léxica opcional)
amor

Grupo
~ + adjetivo

Glosa

Figure 35 The *Ayuda a la redacción* search option

The third advanced search option *Consultas directas* (Direct search) allows finding collocations which are described by a specific LF in DiCE. The user is prompted to introduce a LF and, optionally, a lemma or a LU, and the system retrieves all collocations satisfying the query. For instance, a search for collocations of the lemma *remordmiento* ‘remorse’ described by the LF Sing returns the combinations *acceso de remordimiento* ‘fit of remorse’, *asomo de remordimiento* ‘hint of remorse’ and *punzada de remordimiento* ‘stab of remorse’, see Figure 36. Note that, when selecting a LF, the user has the option of limiting the query only to collocations encoded by the exact LF indicated, or to retrieve all LFs containing the query term. In the latter case, search results also include complex and compound LFs such as Adv$_1$Sing, Sing+Magn, SSingCaus$_1$Manif, etc.



Función: Sing

tipo de combinación:

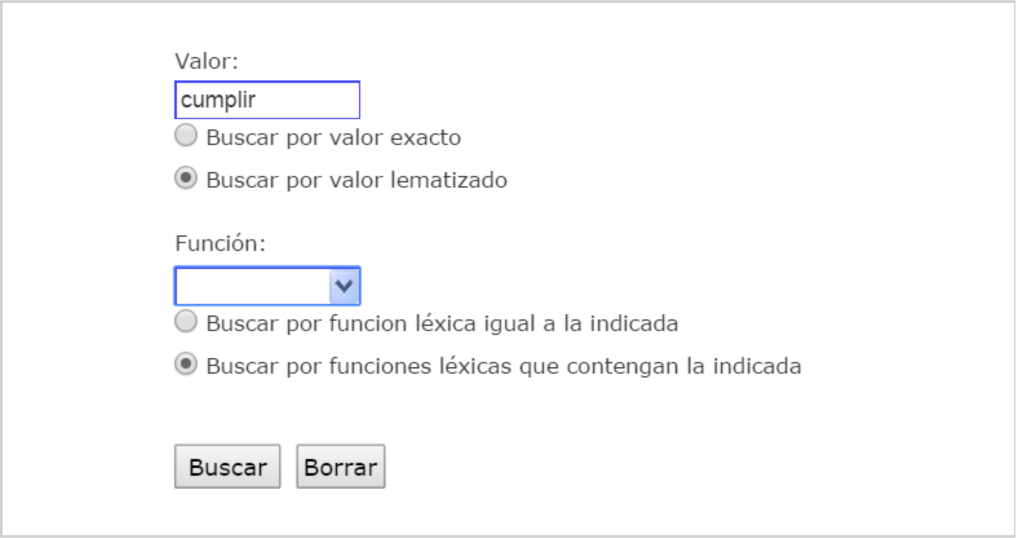
● Buscar por funcion léxica igual a la indicada

○ Buscar por funciones léxicas que contengan la indicada

Lema (unidad léxica opcional): remordimiento

Figure 36 The *Consultas directas* search option

Finally, the *Consultas inversas* (Inverse search) option allows to find collocations containing a specific collocate. Accordingly, the user is asked to introduce a collocate in order to retrieve the bases it can be combined with. Thus, e.g. the query for the collocate *cumplir* 'fulfill' illustrated in Figure 37, returns the bases *deseo* 'wish' and *esperanza* 'expectation', among others. Additionally, the user has the option of limiting the search to a given word form of the collocate (relevant in the case of adjectives) and to a given LF.



Valor: cumplir

○ Buscar por valor exacto

● Buscar por valor lematizado

Función:

○ Buscar por funcion léxica igual a la indicada

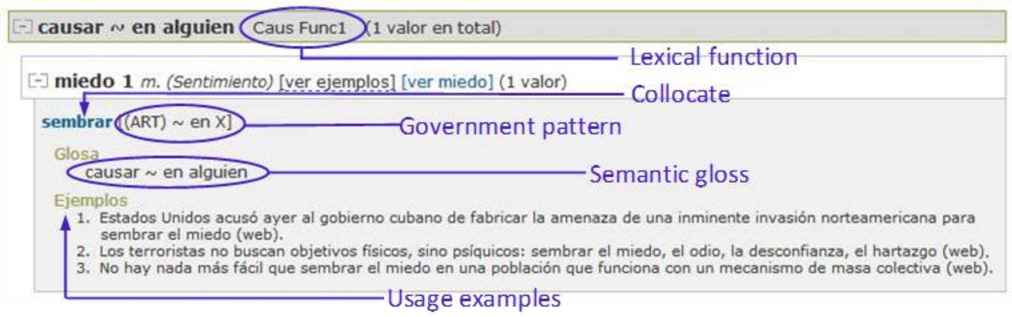
● Buscar por funciones léxicas que contengan la indicada

Figure 37 The *Consultas inversas* search option

In addition to assessing different search options, the usability test also aimed at verifying whether potential users are able to interpret correctly the description of collocations provided by DiCE. A sample lexical entry from the dictionary was already shown in Figure 1 (page 112), nevertheless I provide a brief overview here concerning the information provided on each individual collocation, referring to syntactic pattern, semantic content, government, other grammatical properties and usage examples.

The content of DiCE is limited to combinations involving emotion nouns, and collocations are grouped in five main categories. 1) The first group includes *participant attributes*, i.e. expressions that make reference to a participant of the situation designated by the base. For instance, in the case of the noun *alegría* ‘joy’, we find *loco de alegría* lit. ‘crazy of joy’, describing a person who is feeling a lot of joy. 2) The second group is specified as containing *noun+adjective* collocations, such as *loca alegría* lit. ‘crazy joy’. Note, however, that it also includes other types of combinations where the collocate functions as a modifier of the base, such as *amor a primera vista* ‘love at first sight’. Verbal collocates are found in the categories of 3) *verb+noun* (*dar alegría* lit. ‘give joy’) and 4) *noun+verb* collocations (*la alegría desborda a alguien* lit. ‘joy overflows somebody’), depending on whether the base constitutes the grammatical subject or the complement of the verb. Finally, the last group includes collocations with the pattern 5) *noun+de+noun* (*brote de alegría* lit. ‘a sprout of joy’).

The semantic content of each lexical combination is indicated by a LF, which is translated into a natural language semantic gloss. For instance, in the case of the collocation *sembrar (el) miedo* lit. ‘plant fear’, whose lexical entry is shown in Figure 38, the LF CausFunc$_1$ is paraphrased in the gloss as *causar ~ en alguien* ‘cause in somebody’. Information on the government pattern or other syntactic and morphosyntactic properties is indicated in square brackets. For instance, the collocation entry shown in Figure 38 specifies that the preposition *en* should be used for the second complement of the verb. Finally, corpus examples extracted from language corpora illustrate the use of each collocation.



causar ~ en alguien Caus Func1 (1 valor en total)

miedo 1 *m. (Sentimiento)* [ver ejemplos] [ver miedo] (1 valor)

sembrar [(ART) ~ en X]

Glosa

causar ~ en alguien

Ejemplos

1. Estados Unidos acusó ayer al gobierno cubano de fabricar la amenaza de una inminente invasión norteamericana para sembrar el miedo (web).
2. Los terroristas no buscan objetivos físicos, sino psíquicos: sembrar el miedo, el odio, la desconfianza, el hartazgo (web).
3. No hay nada más fácil que sembrar el miedo en una población que funciona con un mecanismo de masa colectiva (web).

Figure 38 Lexical entry of the collocation *sembrar miedo* ‘plant fear’ in DiCE

5.2.3 Methodology

In the course of the usability experiment participants were asked to query DiCE in order to find the answer to a series of questions. Subsequently, the queries carried out and answers provided by users were evaluated to assess the usability characteristics of the interface. The following subsections describe the methodological aspects of the study in detail.

5.2.3.1 The questionnaire

The questionnaire used in the usability experiment consisted of two main parts, the usability test items proper, which prompted participants to query the DiCE interface, and a post-test aiming to gather information on user satisfaction as well as participants' previous experience with dictionary interfaces.

Usability test items were constituted by 13 questions referring to information related to lexical combinatorics included in the dictionary database. Hence, participants were asked to carry out searches in order to retrieve the answer for each item. Importantly, they were instructed to query the interface even if they were able to provide a solution to a question relying on their own knowledge. Questionnaire items were designed in such a way that, although in most cases they could be resolved through accessing the dictionary content via *Acceso al diccionario*, the search options found in *Consultas Avanzadas* usually provided a more direct path to obtain an answer.

Table 28 shows a few sample questions, together with the optimal query type to be used to retrieve corresponding information. The numbers in the third column indicate the number of questionnaire items included in the usability test which were considered to correspond to the given optimal query type. For instance, in the case of the first questionnaire item, which concerns the meaning of the collocation *reanudar una amistad* 'renew a friendship', it was assumed that the user could most efficiently retrieve the answer through the *¿Qué significa?* search option. When it comes to the next questionnaire item, referring to the LU *cariño* 2, both accessing the list of verbal collocates from the entry of the given LU through *Acceso al diccionario* and using the *Ayuda a la redacción* search option to retrieve the verb+noun and noun+verb collocations of *cariño* 2 included in the dictionary were considered to be efficient search strategies. Note that questionnaire items corresponding to the same optimal query type were not necessarily formulated in the exact same way.

Sample questionnaire item	**Optimal query type**	**Number of corresponding items**
What does *reanudar la amistad* 'renew a friendship' mean?	*Qué significa?*	4
What verbs can be used with the lexical unit *cariño 2* 'affection'?	*Acceso al diccionario/Ayuda a la redacción*	2
Find the adjectives you can use to speak about *amor* 'love' 'that is felt for one another'	*Ayuda a la redacción*	3
Find the collocates of *remordimiento* 'remorse' codified by the lexical function Sing	*Consulta directa*	2
Find all collocations included in the dictionary which contain the collocate verb *cumplir* 'fulfill'.	*Consulta inversa*	2

Table 28 Sample questionnaire items from the usability test

In the case of each of the usability test items, participants were asked to indicate the perceived difficulty of retrieving the answer on a 1–5 Likert-type scale. Following the usability test itself, participants were asked to complete a brief post-test survey in order to measure user satisfaction as well as to gain information concerning participants' previous experience with electronic dictionary interfaces, including *DiCE*. The original questionnaire used in the usability experiment and its English translation are included in Appendix A and Appendix B respectively.

5.2.3.2 Participants

A total number of 26 participants were involved in the usability study. They can be divided into four groups representing different target user-profiles of DiCE. 1) Firstly, eight participants were native Spanish university students. Four of them were undergraduate language majors attending a course called 'Language and technology' at the University of A Coruña, three students were following a one year Master's program in linguistics at the same university, and one participant was an exchange student at the University of Helsinki attending a course in lexicography. 2) The second group was constituted by eight upper intermediate-advanced SFL learners. These participants were university students majoring in Spanish – some of them exchange students – at the University of Helsinki. 3) The third group was made up of five teachers of Spanish or English as a foreign language at the Escuela Oficial de Idiomas in A Coruña. All of them were native speakers of Spanish. 4) Finally, the last group was constituted by five participants who were PhD students of translation studies at the University of Vigo, all native speakers of Spanish. As compared to the other two groups of native speakers, they can be viewed as a group of language professionals, characterized by an increased language awareness and considerable expertise in the use of lexicographic tools.

As it was mentioned above, one of the aims of the post-test survey (see Appendix A and Appendix B), participants were asked to complete after finishing the usability test, was to gather information on their previous experience with DiCE. As their answers revealed, none of the participants had substantial experience with the dictionary interface. Two informants said that they had used it once, another three mentioned having used DiCE, but only during the university course they were taking, and only one informant claimed to have completed the tutorial included in the web site. The remaining 19 participants stated that they had never used the dictionary before, while one participant did not complete the post-test questionnaire.

5.2.4 Procedure

The usability experiment can be divided into three main phases: an informative session, the usability test proper, and the post-test survey.

Previous to the completion of the usability questionnaire, participants received a brief introduction to the concept of collocation, and were given some instructions on the completion of the usability test. They were not, however, instructed in the use of DiCE. Participants belonging to the first two groups (see above), were given all necessary instructions by their course instructors within the framework of a university course they attended, whereas in the case of language teachers and translation students, an informative session was organized to this end.

After having received all necessary information, participants completed the usability questionnaire on their home computers. Given that the aim of the experiment was to assess the dictionary interface from the point of view of a potential user, this was thought to be the best scenario, since participants could explore the dictionary website at their pace, with no time or peer pressure. Participants did not receive any external aid regarding the use of DiCE, therefore, they were left to their own devices regarding whether they preferred to read the introductory user information provided on the dictionary website or to experiment and explore search options themselves. In order to track participants' interactions with the dictionary interface, they were asked to provide the IP address of their computer, together with the time and date of connection, which allowed to retrieve corresponding log files from the DiCE web site.

As it was explained above, participants were given a single questionnaire which contained both the usability test items and the post-test survey. This was distributed and, once completed, returned by participants in electronic format.

5.2.4.1 Data analysis

For the purposes of the quantitative analysis of the results of the usability test the criteria described in Nielsen (1993) were adopted. According to this author, the usability of an interface can be measured along three main aspects: *effectiveness, efficiency* and *user satisfaction*.

Effectiveness of the interaction is measured through the task outcome. In the case of the present study, this aspect of usability was operationalized as participants' performance on the usability questionnaire, which is represented by the number of correct answers provided, i.e. the score obtained by participants.

Efficiency of the interaction, according to Nielsen (1993), is measured through the time and user efforts, i.e. the amount of interaction with the interface, necessary to accomplish the task. In the case of the DiCE usability experiment three parameters were established to measure efficiency: 1) The first of these was the *net time* required to complete the query in the case of each individual test item. 2) The second parameter was constituted by the *effort measure* representing the amount of user interaction involved in attempts to resolve a given questionnaire item. This is calculated as the sum of the number of times a participant chooses, i.e. clicks on, a specific search option, the number of search boxes filled in, the number of search filters set, and the number of times the participant hits the *Search* button before obtaining the definitive answer, i.e. the answer provided on the questionnaire, for the test item. Finally, the third parameter used to measure efficiency was 3) *query-type adequacy*, measured through a score based on the search option used to retrieve a correct answer. A maximum of 3 points were given when participants used the optimal search option for the question with all search filters correctly set; 2 points were given when they used one of the advanced search options – though not the most suitable one –, or when they did not set some of the search filters optimally; and 1 point was given when they used the *Acceso al diccionario* option where an advanced search option would have provided more direct access to the desired information.

While effectiveness and efficiency constitute objective measures, and can be assessed on the basis of participants' answers on the items of the usability questionnaire, together with the data obtained from the log files, the third aspect used to measure the usability of the DiCE interface, *user satisfaction*, is a subjective indicator. It was evaluated on the basis of the results of a post-test survey where participants were asked to report on their experience with the dictionary interface.

As it was explained above, participants' interactions with the dictionary interface were recorded in log files. This information was not only exploited to measure time spent on queries and to calculate the above mentioned effort measure and query-type adequacy scores, but it was also submitted to a detailed qualitative analysis of user actions. As it will be shown, this analysis of participant interactions allowed to detect some specific usability issues presented by the DiCE interface.

5.2.5 Results

This section presents the results of the usability experiment in accordance with the research questions exposed above. Firstly, task effectiveness and efficiency are considered to provide a global measure of how successfully participants queried the DiCE interface, secondly, effectiveness and efficiency scores characterizing queries when retrieving different types of information and corresponding to different groups of questionnaire items are discussed. Thirdly, data relevant to the performance of different participant groups is presented. In relation to the fourth research question, a qualitative analysis of queries using the different search options offered by the DiCE interface is carried out with an aim of pinpointing usability issues. Finally, a brief description of the results of the post-test user survey is offered.

5.2.5.1 Global task effectiveness and efficiency

The mean number of correct answers provided per participant was 9.62 (SD=3.28) out of the total number of 13 test items included in the usability experiment. There were four participants out of the 26 who managed to find the correct answer for all questions and ten participants who answered 11 or 12 questions correctly, whereas two participants only provided one correct answer. Both of the latter abandoned the test without attempting queries for all questionnaire items: one participant carried out queries in order to retrieve the answer for the first three test items, while the other attempted to answer the first seven questions. All other participants carried out queries in an attempt to answer each of the questionnaire items. Figure 39 shows the number of participants per number of correctly answered test items.

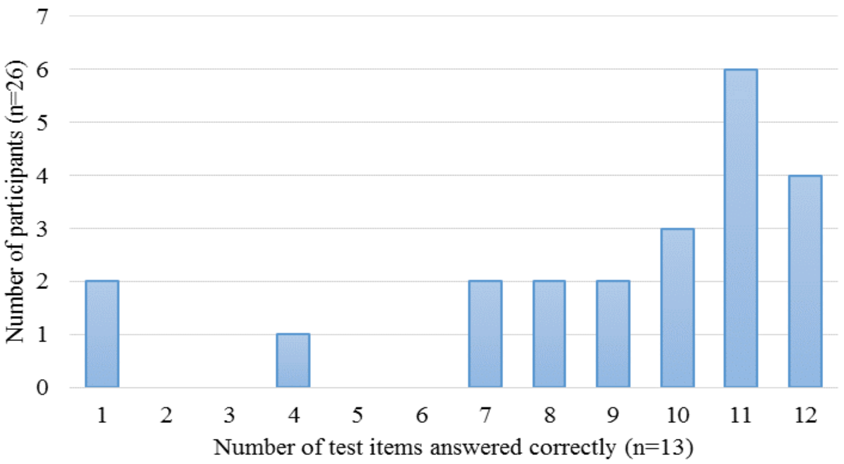


Figure 39 Number of participants per number of correct answers provided

Table 29 shows the mean efficiency scores for participant groups with correct answers for 1-4, 7-9, 10-11, and 12-13 test items, respectively. These scores suggest that participants who obtained twelve or more correct answers, tended to need less time and made less efforts per query, while, at the same time, they often accessed the required information more directly than the rest of the participants. In sum, it appears that participants who obtained the highest scores on the usability test also managed to interact with the dictionary with the most ease.

	Net time (mins)	**Net time per test item[38] (mins)**	**Total efforts**	**Efforts per test item**	**Mean query-type adequacy**
1-4 corr. answ. (n=3)	28.65	3.74	196.67	25.65	2.67
SD	15.13	2.91	136.69	22.91	0.75
7-9 corr. answ. (n=6)	44.97	3.46	264.83	20.37	1.88
SD	18.47	3.45	98.29	20.09	0.97
10-11 corr. answ. (n=9)	49.77	3.83	355.67	27.36	2.35
SD	26.70	4.13	134.17	26.67	0.88
12-13 corr. answ. (n=8)	25.82	1.99	202.63	15.59	2.48
SD	11.18	2.40	52.96	14.13	0.83
Overall mean	**38.86**	**3.16**	**269.27**	**21.74**	**2.32**
Overall SD	**13.12**	**0.81**	**126.66**	**21.99**	**0.90**

Table 29 Summary of overall task efficiency

[38] Note that all means were calculated taking into account the number of test items for which participants attempted to carry out at least one query.

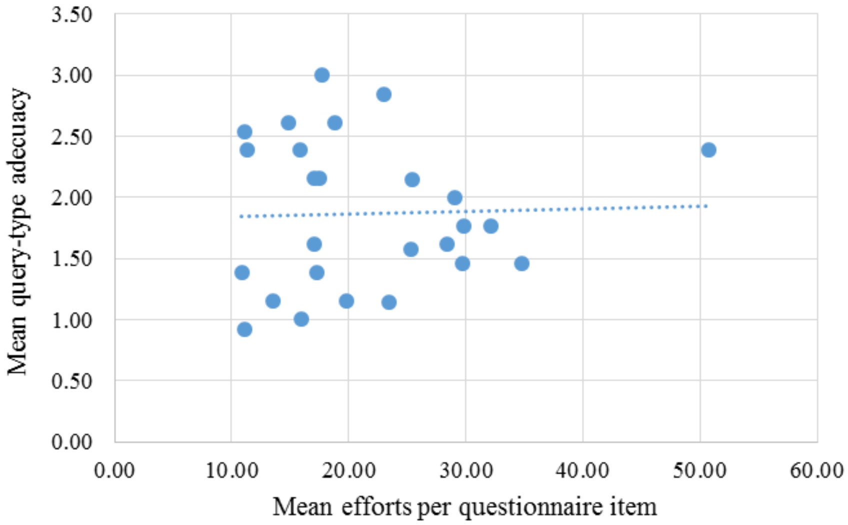


Figure 40 Relationship between mean query-type adequacy score and mean efforts made per questionnaire item

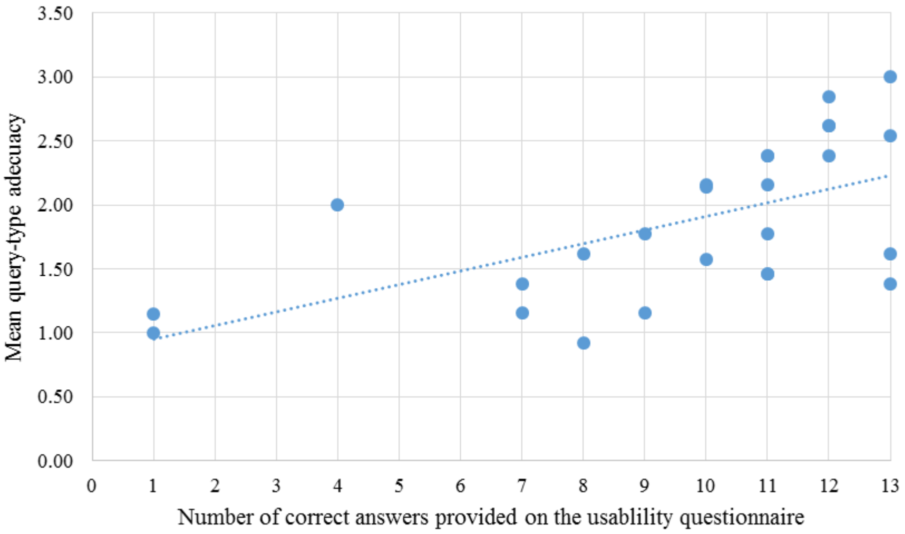


Figure 41 Relationship between mean query-type adequacy score and number of correct answers obtained on the usability questionnaire

Interestingly, as it can be observed on the scatter plot shown in Figure 40, participants with higher query-type adequacy scores did not necessarily make more efforts than those with lower scores. This suggests that, while for some participants the use of the different advanced search options offered by the DiCE was not highly demanding, others did not seem to master their use despite heavy interaction with the interface. At the same time, participants with higher query-type adequacy scores in general tended to provide more correct answers, in contrast with participants who used almost exclusively the more simple, traditional dictionary-type access (i.e. *Acceso al diccionario*) to carry out successful queries. While, this could be expected in the sense that the answer in the case of two particular questionnaire items (Q5 and Q8) could only be obtained using a specific advanced search option, as shown by the scatter plot in Figure 41, the differences in the scores obtained by participants cannot always be explained by these two cases.

Note that the generally high values for mean amount of time and efforts, as well as the low number of correct answers provided are indicative of the difficulty participants had when using DiCE. While this can be in part put down to the relative complexity of the interface, as well as to participants' unfamiliarity with it, the nature of test items could have also intervened. Since the usability test aimed at providing opportunities to manipulate each search option and most search filters, not all test items were formulated in a way that they could be considered to represent look-ups likely to be carried out by the types of users represented by participants. In any case, the overall means of 3.16 minutes and 21.74 efforts required per test item indicate that participants had difficulties in using the interface. In what follows, task effectiveness and efficiency scores are considered according to query type in order to shed more light on the nature of usability problems.

5.2.5.2 Effectiveness and efficiency according to question types

Data regarding the number of participants who managed to find the correct answer in the case of each test item (see Figure 43), or each group of questionnaire items according to the anticipated optimal query type used (see Figure 43), as well as the efficiency scores in the case of each group of questions shown in Table 30 provide information regarding which items of the usability questionnaire were especially problematic.

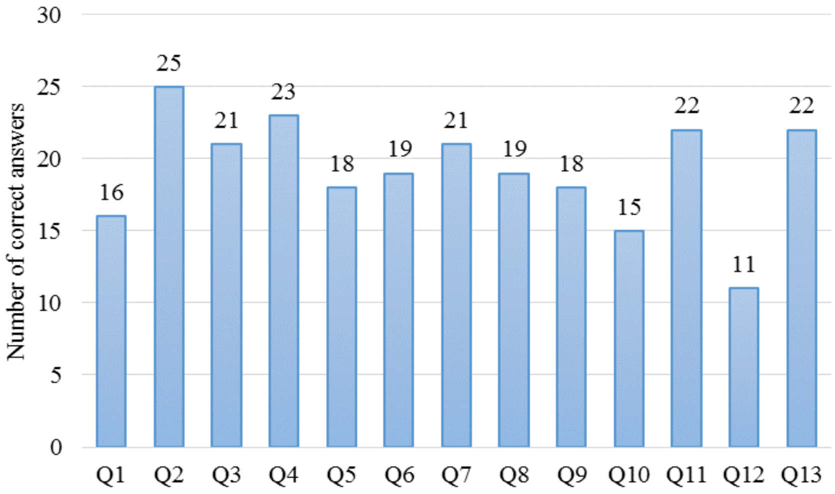


Figure 42 Number of correct answers (n=26) provided per question

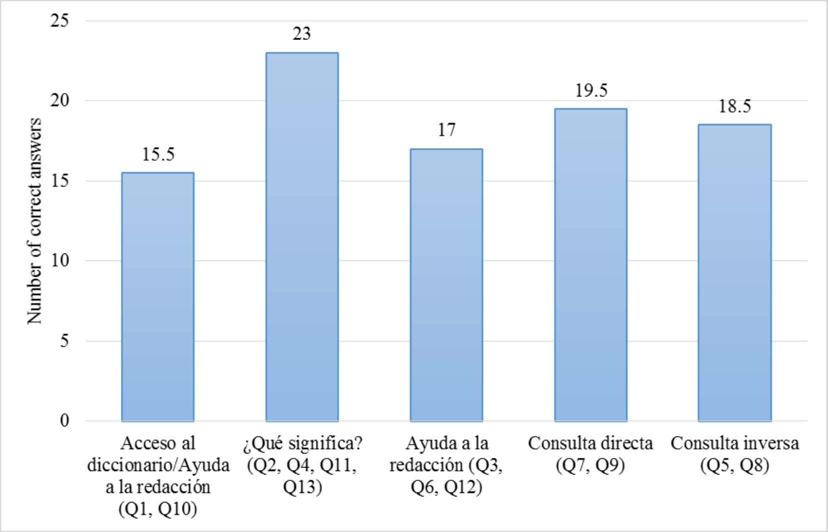


Figure 43 Mean number of correct answers per questions corresponding to an anticipated optimal query type

	Mean correct answers per item	Mean net time per item (mins)	Mean efforts per item	Mean query-type adequacy
Acceso al diccionario / Ayuda a la redacción (Qs 1, 10)	15.50	3.29	17.58	2.94
SD	0.50	3.01	16.67	0.25
Qué significa? (Qs 2, 4, 11, 13)	23.33	2.13	18.15	2.24
SD	1.22	2.88	22.56	0.96
Ayuda a la redacción (Qs 3, 6, 12)	17.00	3.08	20.07	1.82
SD	4.32	2.96	18.83	0.98
Consulta directa (Qs 7, 9)	19.50	3.62	26.29	2.23
SD	1.50	4.23	22.48	0.83
Consulta inversa (Qs 5, 8)	18.50	4.62	31.27	2.78
SD	0.50	4.29	25.63	0.47

Table 30 Summary of effectiveness and efficiency for question groups according to optimal query type

With a mean of 23.3 correct answers per question, participants were most successful in answering questionnaire items categorized as most suitably queried using the *¿Qué significa?* search option (questions 2, 4, 11 and 13). Data indicate that in the case of these test items, participants also spent less time on the queries, and made on average less effort than in the case of other questionnaire items – except for questions 1 and 10. These scores indicating higher effectiveness and efficiency might be explained by that the corresponding test items required participants to search for the lexical entry of a given item, in this case, of a specific collocation, which is precisely why they would do in a prototypical dictionary look-up. Furthermore, it should be also noted that the wording of questions 2 and 4 made explicit reference to the *meaning* of a collocation, which might have served as a clue for participants to try and use the *¿Qué significa?* (What does it mean?) advanced search option. Nevertheless, the query-type adequacy score, shown in Table 30, indicates that in fact a number of participants did not make use of this query type, but retrieved the desired information through *Acceso al diccionario* instead.

Interestingly, the second highest mean number of correct answers (19.5) was obtained in the case of items which were classified as optimally queried using *Consulta directa* (questions 7 and 9), despite the fact that these involved the use of LFs – which participants were not expected to be familiar with. As Table 30 shows, in the case of these questions, the mean time spent on the queries (3.62 minutes per item) as well as the mean number of efforts made per query (26.29) was the second highest. The query-type adequacy score again shows that instead of making use of the optimal search option, some participants used *Acceso al diccionario*. At the same time, in the case of question 9, 10 participants did not set the search parameters completely (see 5.2.5.4.B).

A slightly lower number of correct answers were obtained (a mean number of 18.5) in the case of questions 5 and 8, which prompted finding collocations in DiCE starting from the collocate, and could be resolved using *Consulta inversa*. Note that, as mentioned earlier, these were the only questionnaire items where participants necessarily had to make use of a specific advanced search option to retrieve the required answer, while in the case of all other items, the required information could be obtained through using *Acceso al diccionario*. This explains why participants spent the highest mean time (4.62 minutes per item) and made the most efforts, a mean of 31.27 in the case of these queries, while the mean query-type adequacy score is the highest (2.78), since, with the exception of one, participants who provided correct answers for these questionnaire items all used *Consulta inversa*.

In the case of the test items where subjects were expected to use *Ayuda a la redacción*, collocations had to be searched for starting from the base and the meaning of the combination. While questions 3 and 6 were answered correctly by a relatively high number of participants (21 and 19, respectively), question 12 proved more problematic, since here participants were also asked to identify a specific LU based on an example sentence. Given this additional difficulty, correct answers were obtained only by 11 participants. Furthermore, the low query-type adequacy score shows that more than half of the subjects did not make use of the optimal advanced search option, but used *Acceso al diccionario.*

Finally, in the case of questionnaire items 1 and 10, participants were asked to list collocates of a specific LU corresponding to a given syntactic pattern, a query for which both *Acceso al diccionario* and *Ayuda a la redacción* were considered to constitute optimal access paths. These two questionnaire items are among the three questions with lowest number of correct answers. In the case of question 1, having retrieved the lexical entry of the base noun, three participants mistook the semantic gloss for the target collocate (for more detail see 5.2.5.4.A), while another four participants did not access the list of collocations; instead, they aimed to deduce possible collocates from the example sentences displayed in the initial view of the entry of the base LU. Much of this low performance can probably be attributed to the greater unfamiliarity with the DiCE interface at the beginning of the usability test session[39]. In the case of question 10, similarly to the case of question 12 described above, the low performance of participants was mainly due to their failure to identify the correct LU in the example sentence provided. While the mean net time and efforts are in the mid-range in the case of this set of two questionnaire items, the query type adequacy is high, which is because the use of *Acceso al diccionario*, the option most preferred by subjects, was scored as one of the optimal access routes.

Table 31 provides a summary of the search options used by participants to obtain correct answers in the case of each questionnaire item. The highlighted squares represent the optimal query type in each case, which, as we can see, for the majority of questionnaire items was the most frequently used search option. However, it can also be seen that, overall, most correct answers provided in the usability experiment were obtained through the default search option, *Acceso al diccionario.*

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	TOTAL
Acceso al diccionario	**10**	8	9	7		12	5		5	**12**	6	9	12	95
¿Qué significa?		**17**		**16**	1						**14**		**10**	58
Ayuda a la redacción	**6**		**12**			**7**				**3**	1	**2**		31
Consulta directa							**16**		**13**					29
Consulta inversa					**17**			**19**			1			37

Table 31 Summary of the number of correct answers generated using each query type per question

[39] Participants of the usability experiment carried out the queries following the order of questionnaire items, and, only four of them were observed to go back to revise their answers or query a question that was initially left unanswered.

5.2.5.3 Effectiveness and efficiency according to user profiles

As it was discussed in 5.2.3.2, participants of the usability experiment were considered to represent four main user groups: Non-native speaker university students, native speaker university students, native speaker foreign language teachers and native speaker translators. Table 32 shows the summary of mean effectiveness and efficiency scores, while Figure 44 indicates the mean number of correct answers provided by participants according to user profiles. As a general observation, note that, as shown by the standard deviation scores (SD), considerable individual differences could be observed in each participant group regarding the use of the DiCE interface.

	Mean number of correct answers	**Net time (mins)**	**Mean net time per test item (mins)**	**Mean total efforts**	**Mean efforts per test item**	**Mean query type adequacy**
Non-native students (n=8)	10.00	47.36	3.87	309.88	25.30	2.47
SD	3.87	28.50	4.33	165.78	26.91	0.84
Native students (n=8)	10.38	35.68	2.74	240.00	18.46	2.00
SD	1.87	20.51	3.49	80.61	18.75	0.96
Native teachers (n=5)	6.80	40.27	3.66	302.80	27.53	3.15
SD	3.76	17.65	2.94	128.60	23.48	2.06
Native translators (n=5)	10.60	28.91	2.22	217.60	16.74	2.38
SD	1.50	10.32	1.80	68.18	13.57	0.87

Table 32 Summary of efficiency and efficacy scores according to participant groups corresponding to different user profiles

As it can be seen both in Table 32 and Figure 44, the group of translation students scored highest in the test, since they obtained the highest mean number of correct answers (mean=10.6, SD=1.50). Native Spanish university students performed slightly better concerning the number of correct answers (mean=10.38, SD=1.87) than non-native university students (mean=10.0, SD=3.87), while the group of native Spanish foreign language teachers seems to have had the most difficulties using the *DiCE* interface, in that they provided the lowest mean number of correct answers (mean= 6.8, SD=3.76).

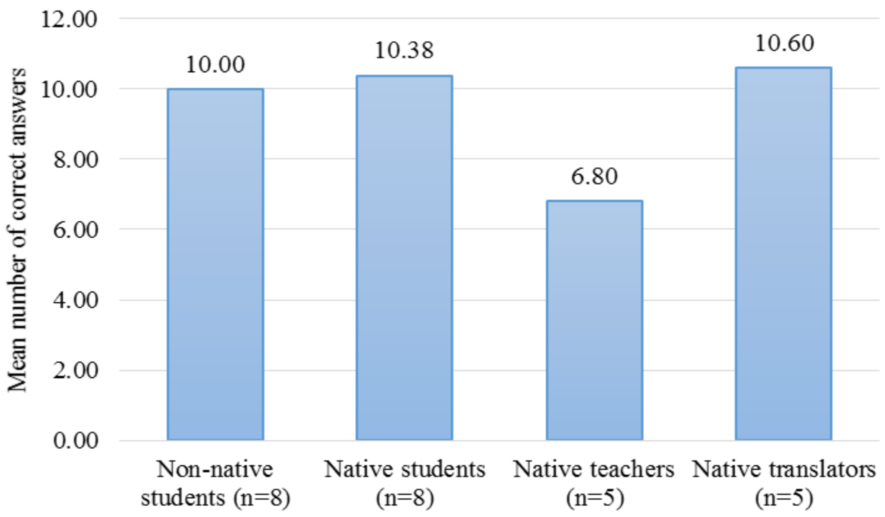


Figure 44 Mean number of correct answers according to participant groups corresponding to different user profiles

With respect to efficiency measures, the best results can again be seen in the case of the group of native speaker translation students, who needed the least mean amount of time (mean=2.22 minutes, SD=1.80 minutes) and made the least mean number of efforts (mean=16.74, SD=13.57) per questionnaire item. Since these scores can be interpreted as indicating the ease of use of the interface, it may be concluded that, as it was expected, participants belonging to this group, assumedly having more experience in using different language resources or tools, were also better at manipulating DiCE participants in the other three groups. When it comes to comparing native and non-native university students, who performed similarly when it comes to the mean number of correct answers, we can see that native students seemed to use the dictionary interface with more apparent ease, as they applied less time (mean=2.74 minutes, SD=3.49 minutes) and made less efforts (mean=18.46, SD=18.75) per item in completing the test than their non-native peers (mean net time per item=3.87 minutes, SD=4.33 minutes; mean efforts per item=25.30, SD=26.91). Finally, the group of foreign language teachers, who obtained the lowest mean number of correct answers in the usability test, spent a slightly lower mean amount of time (mean=3.66 minutes, SD=2.94 minutes) and applied somewhat more efforts (mean=27.53, SD=23.48) per questionnaire item than the previous group. For data concerning mean net time and the mean number of total efforts per item see also Figure 45 and Figure 46. Although the differences observed in the performance of participant groups – in part – seem to comply with expectations, it has to be noted that according to the results of the Kruskal-Wallis test the null hypothesis cannot be rejected in the case of neither the test scores, i.e. number of correct answers (χ^2=3.978, p=0.264), nor the average time spent on queries per questionnaire item (χ^2=3.996, p=0.262), or the average efforts made per item (χ^2=4.646, p=0.2). Therefore, there is not enough evidence to claim that the differences observed are statistically significant.

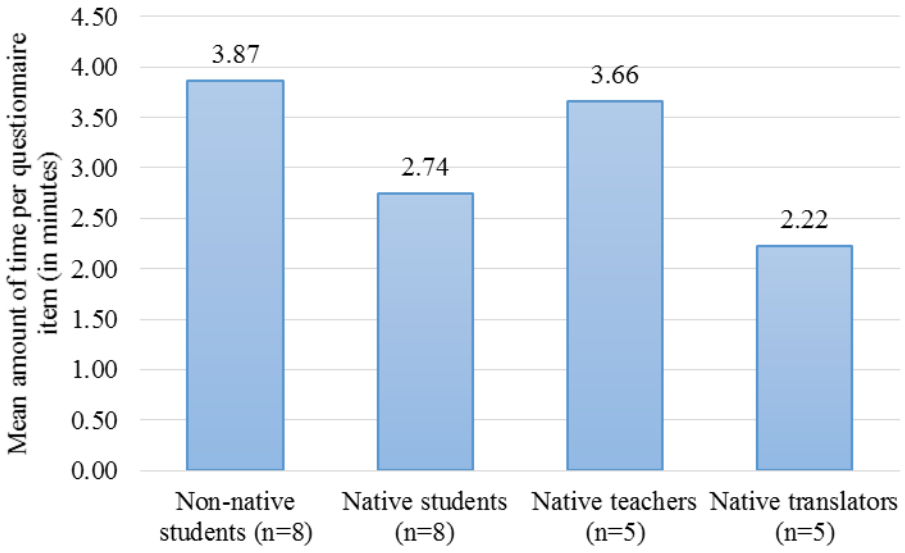


Figure 45 Mean net time (in minutes) per questionnaire item according to participant groups

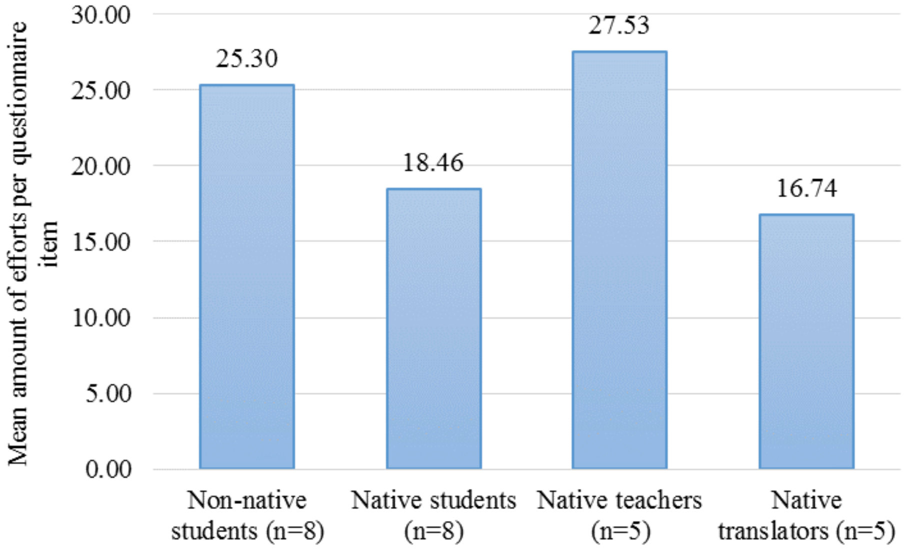


Figure 46 Mean number of efforts made per questionnaire item according to participant groups

5.2.5.4 Usability of different search options

This section considers participant interactions with each of the search options offered by the DiCE interface. As we have seen, the items of the usability questionnaire were designed in a way that they encourage users to experiment with the different advanced search options available in the DiCE web interface. Figure 47 shows the number of correct answers generated using each search option. The numbers in brackets represent the overall number of test items where each of these was considered to constitute the optimal query type. As it was indicated earlier, information required by all but two test items could be retrieved through *Acceso al diccionario*, therefore, in the case of this search option, the second number in the bracket represents the overall number of test items where this module could be successfully used. In the case of *Ayuda a la redacción*, we have seen that, in addition to the questionnaire items which were considered to be optimally queried using this option, with another two items both *Acceso al diccionario* and *Ayuda a la redacción* were thought to be equally efficient options (see Table 28), hence the second number in the bracket represents the sum of all of these items.

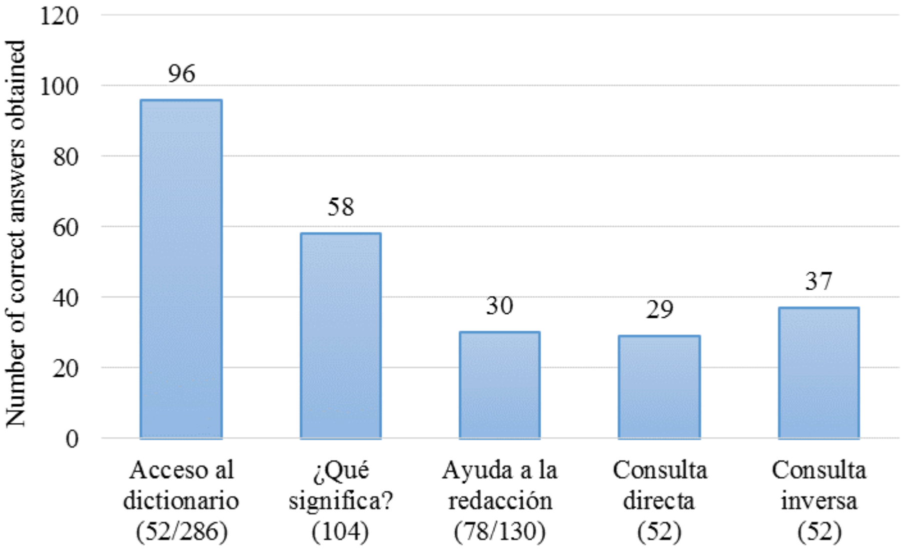


Figure 47 Number of correct answers generated per search option

Figure 48 shows the number of participants who used each search option at least once to carry out a query, together with the number of participants who used the given option successfully at least once.

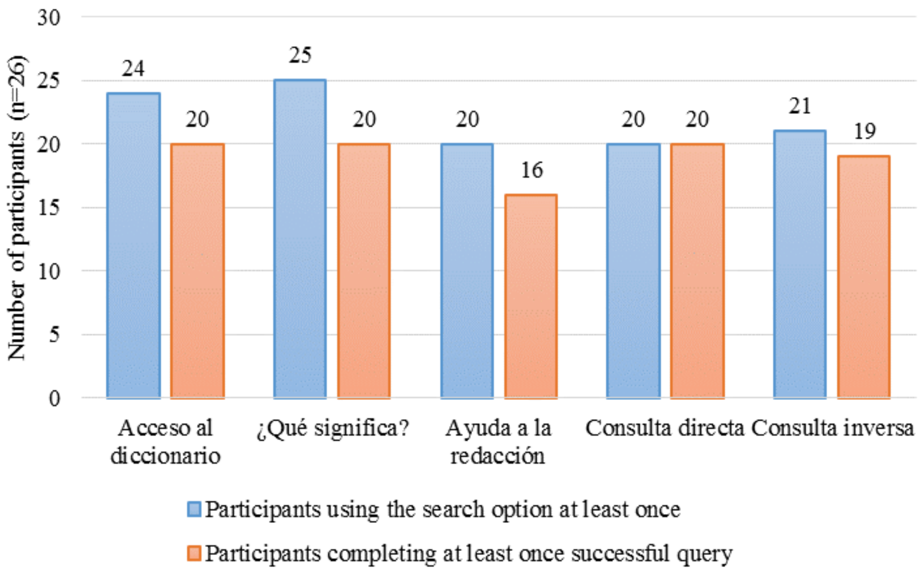


Figure 48 Number participants carrying out successful and unsuccessful queries per search option

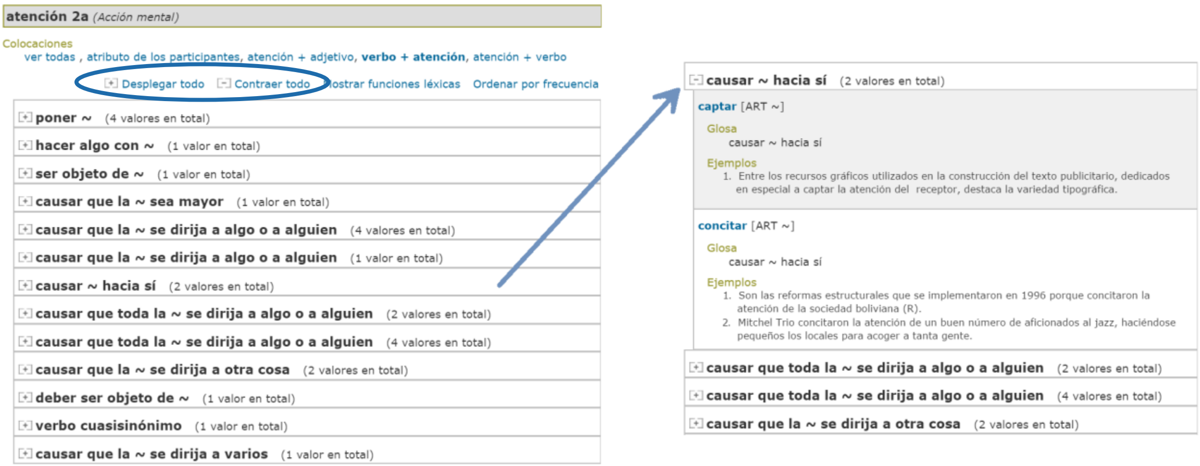
As the results presented above suggest, participants used *Acceso al diccionario* most often, which is not surprising, since it constitutes the default type of access offered by the web interface, furthermore, it presents lexical information in a way which is more similar to the dictionaries participants may have been familiar with. Among the four advanced search options *¿Qué significa?* was the most often used query type, while *Consulta inversa*, *Ayuda a la redacción* and *Consulta directa* were less frequently used by participants. In the following sections I provide an analysis of the use of each search option, taking into account qualitative data obtained from the log files, with the aim of exploring what features of the DiCE interface resulted particularly problematic from a usability point of view.

A. *Acceso al diccionario*

As shown in Figure 48, *Acceso al diccionario* was used at least once by 24 participants, 20 of whom managed to use it successfully. Note that the two participants who did not use this search option at all, as well as another three participants who did not manage to use it successfully, made use of the advanced search options instead to carry out successful queries.

As explained above, similarly to collocation dictionaries in general, *Acceso al diccionario* provides access to collocations through the lexical entry of the base. An issue specifically related to this search option was that participants often introduced a collocate in the search box, and, consequently, failed to obtain any search results.

As it was already explained in 3.4.2.1.B, similarly to paper collocation dictionaries, DiCE presents collocations within the lexical entry of the base grouped according to their syntactic pattern and their meaning, indicated by a semantic gloss. While users generally had no difficulty in identifying the appropriate collocate group in terms of syntactic pattern, semantic glosses did lead to certain confusion. Currently, when accessing a collocate group (e.g. verb+noun) from the main entry of the base, the user is provided with a list of semantic glosses, aimed at facilitating navigation in cases when a large number of collocates are available. In order for the interface to display the list of collocates, the user has to click on the tab of the corresponding gloss (see Figure 49).



atención 2a *(Acción mental)*

Colocaciones

ver todas , atributo de los participantes, atención + adjetivo, **verbo + atención**, atención + verbo

⊞ Desplegar todo ⊟ Contraer todo Mostrar funciones léxicas Ordenar por frecuencia

⊞ **poner ~** (4 valores en total)

⊞ **hacer algo con ~** (1 valor en total)

⊞ **ser objeto de ~** (1 valor en total)

⊞ **causar que la ~ sea mayor** (1 valor en total)

⊞ **causar que la ~ se dirija a algo o a alguien** (4 valores en total)

⊞ **causar que la ~ se dirija a algo o a alguien** (1 valor en total)

⊞ **causar ~ hacia sí** (2 valores en total)

⊞ **causar que toda la ~ se dirija a algo o a alguien** (2 valores en total)

⊞ **causar que toda la ~ se dirija a algo o a alguien** (4 valores en total)

⊞ **causar que la ~ se dirija a otra cosa** (2 valores en total)

⊞ **deber ser objeto de ~** (1 valor en total)

⊞ **verbo cuasisinónimo** (1 valor en total)

⊞ **causar que la ~ se dirija a varios** (1 valor en total)

⊟ **causar ~ hacia sí** (2 valores en total)

captar [ART ~]

Glosa

causar ~ hacia sí

Ejemplos

1. Entre los recursos gráficos utilizados en la construcción del texto publicitario, dedicados en especial a captar la atención del receptor, destaca la variedad tipográfica.

concitar [ART ~]

Glosa

causar ~ hacia sí

Ejemplos

1. Son las reformas estructurales que se implementaron en 1996 porque concitaron la atención de la sociedad boliviana (R).
2. Mitchel Trio concitaron la atención de un buen número de aficionados al jazz, haciéndose pequeños los locales para acoger a tanta gente.

⊞ **causar que toda la ~ se dirija a algo o a alguien** (2 valores en total)

⊞ **causar que toda la ~ se dirija a algo o a alguien** (4 valores en total)

⊞ **causar que la ~ se dirija a otra cosa** (2 valores en total)

Figure 49 Opening a gloss tab in a DiCE lexical entry

Some of the participants appeared to misinterpret glosses and mistook them for collocates, as suggested by their answers in the usability test where they enumerated both types of elements as representing the combinatorics of the base. A few participants did not seem to be aware of that in order for the interface to display collocations, it is necessary to click on the desired gloss first. As a consequence, there participants merely provided a list of glosses in the test, instead of the collocates they were asked to enumerate. Furthermore, it was also observed that semantic glosses did not appear to be particularly useful in facilitating browsing through combinatory information, since, especially in the case of verbal collocates (e.g. *captar/centrar atención* 'catch/focus attention', *sembrar miedo* 'instill fear'), participants often had to open tabs corresponding to a series of different glosses before they managed to find the desired collocation.

It was also noted that the buttons *Desplegar todo* (expand all) and *Contraer todo* (contract all), located on the top of each screen displaying search results (see above, in

Figure 49), were used by a very low number of participants. These two buttons are aimed to facilitate navigation through the dictionary entry. Clicking on *Desplegar todo* allows to display the full list of collocations collapsed under the corresponding semantic glosses on the initial screen. Therefore, it is especially useful when the user is unsure of which gloss to choose, and it would have been helpful for participants of the usability test. On the contrary, the button *Contraer todo* serves to hide collocates, allowing the user to look through the available glosses at a glance.

B. Consultas avanzadas

¿Qué significa?

As we have already seen, *¿Qué significa?* was the most frequently used advanced search option in the usability test. All but one participant used it at least once to carry out a query, and 20 participants managed to use it successfully (see Figure 48). In what follows I discuss the two main difficulties encountered in relation with this query type.

Firstly, probably due to lack of familiarity with the notion of collocation and the terminology applied in DiCE, participants had difficulty in using the search form, where the base and the collocate have to be introduced separately in individual search boxes (see above, in Figure 34). For instance, one participant typed the whole target collocation in the box corresponding to the collocate, while a number of participants confused the two elements of the collocation when introducing them in the search boxes. In addition, some participants were observed to confuse the collocate with the semantic gloss of a collocation, as they typed a gloss in the search box corresponding to the collocate.

Secondly, the majority of participants were observed to click on the search button at least once without having specified both elements of the collocation, which is obligatory in the case of this search option. This can be explained by that these participants did not perceive which fields had to be obligatorily filled in for the query. While, another feasible explanation is that the limited number of searches performed during the usability test was not enough for participants to learn to distinguish between the available search options, thus they might have attempted to perform queries regarding the combinatorics of a given base or a collocate using *Ayuda a la redacción* instead of *Acceso al diccionario* or *Consultas inversas*.

Ayuda a la redacción

One of the less frequently used advanced search options was *Ayuda a la redacción*, as shown by both the number of correct answers provided and the number of participants who used this query type at least once (see Figure 48).

This search option requires introducing a base, a syntactic pattern and semantic gloss to find corresponding collocations. As in the case of the *¿Qué significa?* option, the most common problem was related to participants confusing the elements of the collocation, namely typing a collocate in the search box corresponding to the base. Another issue concerned users failing to specify the syntactic pattern of the collocation, which is obligatorily required in the query. Note, however, that when they did so, participants did not seem to have difficulties in determining the syntactic pattern of the target collocation.

Consulta directa

As it can be seen in Figure 48, all 20 participants who used the *Consulta directa* option at least once managed to complete a successful query, despite the fact that this search option involves the use of LFs, a formal tool most subjects were not familiar with.

As explained in Section 5.2.2, in order to carry out a query using *Consulta directa*, users have to specify a LF and, optionally, a base. When introducing a LF, one can choose to search for collocations described by a combination of LFs containing the LF introduced as search term or by the exact same LF, the former being the default option. One of the issues encountered was related to precisely this distinction. For instance, when asked to search for collocations of the noun *remordimiento* 'remorse' codified by the simple LF Sing (test item 7), a number of participants used the default option, and thus obtained collocations described both by the simple LF Sing and by the complex LF SSingCaus$_1$Manif. In several cases, participants provided all combinations displayed in the search results on the answer sheet, instead of selecting the ones described by Sing. When scoring the tests, I opted for accepting this solution, because it could, be interpreted as either resulting from participants' unfamiliarity with LFs or from the ambiguity in the wording of the questionnaire item. Nevertheless, the fact that participants frequently used the default query settings, might also mean that the option of limiting the query to an exact LF is not salient enough.

Another issue observed in the case of *Consulta directa* involved queries targeting complex LFs or configurations. Question 9 of the usability test required participants to search for the configuration of LFs Magn+A_1Manif. This type of LF has to be introduced in the search form through selecting each of its elements one by one from a drop-down menu. This appeared to cause difficulties for a number of participants, some of whom carried out searches for a single element of the LF instead, typically Magn or Manif. Note that, in these cases the results obtained through the default query setting, discussed above, included the desired information, hence participants were able to provide a valid answer in the test.

Consulta inversa

Consulta inversa was used at least once by 20 participants, of whom 19 managed to carry out at least one successful query (see Figure 48). As it was already indicated, questionnaire items intended to test the usability of this search option differed from the rest in that the required information could not be retrieved through using the default *Acceso al diccionario* option.

As it was explained in 5.2.2, the *Consulta inversa* search option serves to find the bases with which a given collocate can be combined. In order to launch a query it is necessary to introduce the collocate and, optionally, a LF. In the case of the collocate, users can choose to search for a word form (this is relevant in the case of the feminine or the masculine form of an adjective) or the lemma, the first one being the default option.

Question 8 of the usability test required subjects to retrieve nouns co-occurring with the collocate adjective *negro/a* 'black'. A closer look at the queries revealed that a number of participants did not manage to search for the lemma directly, but carried out two queries instead, one involving the feminine and the other the masculine form of the adjective. In fact, some of the user interactions suggested that participants may not have been familiar with the distinction between lemma and word form. For instance, in certain cases, after having received no results for a specific word form, participants carried out a lemma search with the same search term.

Finally, it was also observed that participants in general tended to confuse the *Consulta inversa* search option with *Consulta directa*, which – in part – might be a result of their lack of familiarity with the terminology related to collocations used in DiCE. Although the two search options have essentially different functions, at a glance they may appear to be similar, as they both contain a search box which requires the user to introduce one of the elements of a collocation – the base in the case of *Consulta directa* and the collocate in the case of *Consulta inversa*. Nevertheless, as it has been noted, the confusion of the different query options might have resulted from participants' general lack of familiarity with the DiCE interface.

5.2.5.5 User satisfaction

As it was explained earlier, information on user satisfaction was collected in a post-test questionnaire (see Appendix A and Appendix B). In addition, as part of the usability questionnaire, following the completion of each test item, participants were asked to assess its difficulty on a 1–5 Likert-type scale.

When it comes to participants' opinion regarding the DiCE interface, I will consider the answers provided on two of the survey questions here. In the case of the first question, which inquired whether they would use the dictionary in the future, 20 out of the 25 participants who completed the survey replied "yes", while the remaining five said they "may" use it, see Figure 50. The second survey item was concerned with whether participants would recommend DiCE to other potential users. 19 subjects replied "yes", three said that they "may" recommend it, and the remaining three participants said that they would recommend it, although with initial practice to facilitate use, or they would recommend only the simpler features, see Figure 51.

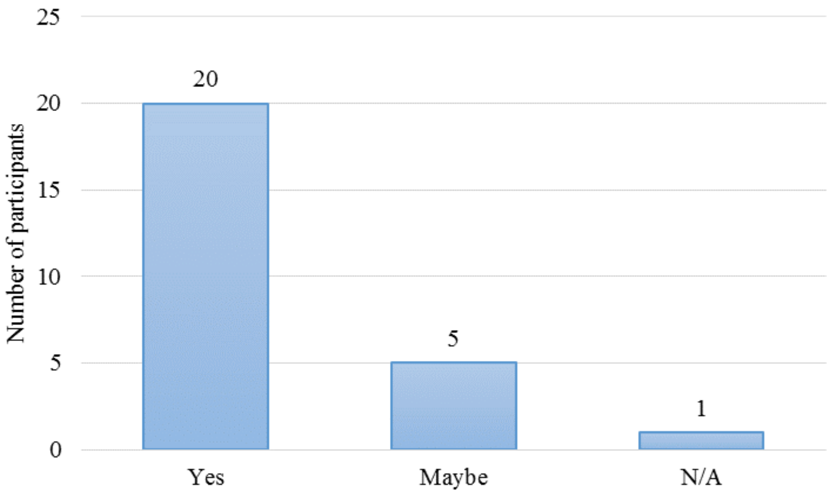


Figure 50 Summary of answers to the survey question "Will you use DiCE again?" (n=26)

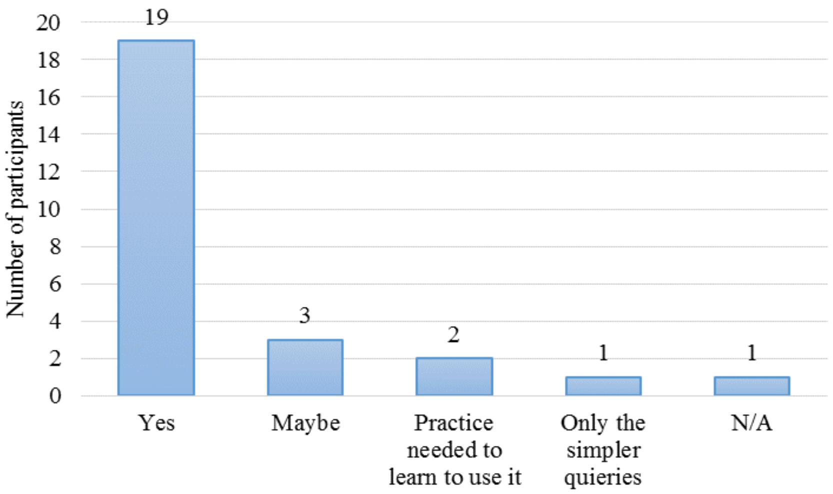


Figure 51 Summary of answers to survey question "Would you recommend DiCE to others?" (n=26)

In sum, participants' answers seemed to reveal a clearly positive attitude towards DiCE, with some reservations concerning the complexity of the interface. This last point is also apparent if we observe the difficulty score assigned to questionnaire items after having attempted to carry out the required queries. The mean difficulty score given by participants was 2.65 (SD=0.77) on a 1-5 scale, which suggests that they found querying the interface relatively complicated.

5.2.6 Discussion

The study conducted to assess the usability of the DiCE interface addressed five main research questions. The present section summarizes the results presented above in relation to these, while it also considers their implications.

The first research question referred to global task effectiveness and efficiency characterizing participants' interactions with the dictionary interface when aiming to retrieve information required by the test. Effectiveness was measured through the number of correct answers provided on the usability questionnaire. The mean number of correctly answered test items per participant was 9.62, with 14 out of 26 participants providing correct answers on at least eleven out of the total number of thirteen test items. In comparison, three participants managed to answer only at most four questions correctly. These results might be considered unsatisfactory, since they seem to indicate that one out of each four test items, referring to a piece of information which is in fact available in the dictionary database, remained unanswered or resulted in an incorrect or incomplete answer. As it was observed participants' relatively poor performance can be attributed in part to the complex nature of the user interface of the dictionary, and in part to their unfamiliarity with the DiCE and, possibly, collocation dictionaries in general. However, when assessing the outcomes of the experiment, it should also be taken into account that at least a number of questions constituting the usability test were themselves rather complex, and on occasions required searching for information that is in fact irrelevant for the general dictionary user, e.g. the LF describing a given collocation.

Task efficiency was represented by three different measures: time spent on queries, an effort measure, calculated on the basis of user actions, and query type adequacy. While participants needed on average a rather long time (38.86 minutes), and, accordingly, a great amount of interaction with the interface to complete the usability questionnaire, it was noted that those making a higher amount of efforts did not necessarily seem to have managed to master the advanced query types and hence obtain higher query-type adequacy scores. At the same time, participants with the highest number of correct answers appeared to have gained mastery of the DiCE interface with most ease, as shown by the fact that they completed the test in less time, through interacting more efficiently with the interface, and making use of the optimal search options more often. The results also hint at the existence of considerable individual differences between potential dictionary users both in terms of their skills in using an online tool, and their willingness to learn to use it.

The second research question concerned effectiveness and efficiency scores according to types of test questions, in other words, the types of information participants were required to retrieve from the dictionary. Data showed that the questions with highest number of correct answers were those that required participants to search for a given collocation, optimally using the *¿Qué significa?* search option. It was suggested that a plausible explanation for this may be that the task involved is closest to prototypical dictionary look-ups, i.e. searching for a given lexical item to obtain information related to it. The remaining test items were characterized by lower rate of correct answers, and, in general, lower efficiency, i.e. more time and more efforts applied in queries. Another observation referred to the fact that most correct answers provided on the usability test were obtained through the *Acceso al diccionario* search option. This suggests that at least novel users prefer to make use of this default access route, which at the same time resembles traditional dictionary interfaces most closely. It was also noted that participants, in general were able to make use of alternative – more advanced – search options, when necessary. All these results point in the direction that, on the one hand, it might not be desirable to split different query options so drastically as it is done in DiCE, and that the implementation of an all-purpose search field, or at least more versatile search options might be desirable, if the goal is to make different access possibilities offered by the dictionary to be more salient and better exploited by users. On the other hand, users, who appear to be biased by the types of look-ups generally possible in dictionaries, could benefit from learning about alternative access paths and the full spectrum of information they can obtain from a dictionary.

The third research question dealt with the performance of participants according to different user profiles. The results showed that the group of native speaker translation students performed best, which is not surprising, since participants in this group were probably more used to dealing with different lexical tools. In contrast, the poorest results – in terms of test scores and efforts made – were obtained by the participant group constituted by language teachers. In their case, it should be noted that they belonged to an older age group than the rest of the participants, who were undergraduate or graduate students at different universities, consequently, they could have had less experience or skill in using web interfaces in general. Regardless, this finding is somewhat alarming in the sense that, even though DiCE in its current state does not constitute a prototypical tool for language learners, professionals in foreign language teaching should be expected to be ready to use different kinds of potential learning resources.

The forth research question referred to exploring specific usability problems affecting the DiCE interface. In order for this, log files which recorded user interactions with the dictionary were submitted to a qualitative analysis. Most problem areas identified and described in the previous section resulted from informants’ difficulties in interpreting the dictionary content and the presentation of lexicographic data. An important problem observed concerned the lack of participants' familiarity with the notion of collocation and related terminology applied in DiCE. Participants’ tendency to confuse the elements of a collocation lead to difficulties in using a number of search options. For instance, as it was explained above, in the case of the *¿Qué significa?* search option, a number of participants were observed to confuse the base and the collocate when required to introduce them in individual search boxes, and, as a consequence, they did not obtain the desired search result. Another problem was constituted by the confusion of collocates with semantic glosses, the latter serving as indications of the approximate meaning of collocations in DiCE. As mentioned above, when accessing collocation entries through *Acceso al diccionario,* some participants included glosses in their answers on the usability test as if they were collocates. In addition to the terminology related specifically to collocations, some participants seemed to be unfamiliar with the more general concepts of word form and lemma, which, as it was explained above, could be observed in the case of queries carried out using *Consulta inversa.* A further issue was constituted by participants' difficulties in distinguishing word senses, i.e. LUs. In fact, the two questionnaire items for which the lowest number of correct answers was obtained involved identifying a particular LU on the basis of example sentences.

What follows from the above observations is that the DiCE interface can be improved on a number of features so that it become more user-friendly, while the results of the usability experiment also emphasize the importance of users' reference skills and familiarization with the dictionary. Possible changes to the current design that could considerably improve the usability of DiCE include providing a more consistent exemplification of the content to be introduced in each search box, giving a clear indication of which search boxes have to be filled in obligatorily and of how search filters should be used, as well as the enhancement of the visibility and distinguishability of navigation aids, e.g. semantic glosses and buttons that allow to expand and contract information shown on the screen. As for a more comprehensive reorganization of the web interface or the design of a new learning tool, it might be useful to incorporate an all-purpose search box, in which users could introduce a base, a collocate or a whole collocation to retrieve corresponding information. Classification and searching of collocations according to semantic glosses is an issue that deserves special attention. As the results of the usability experiment showed, participants had certain difficulty in interpreting glosses, it is not clear, however, whether this was due to their unfamiliarity with them or rather to the lack of transparence in the formulation and presentation of the glosses themselves. Regardless, as it was mentioned in 3.4.2.1.B, semantic classification and description of collocations constitutes one of the main organizing principles in combinatory dictionaries, and, as emphasized in 3.4.2.1.F, it should certainly be better exploited for promoting the ease of navigation in the dictionary as well as for enhancing language learners' collocation production. These usability issues will be revisited in relation to the collocation learning tool proposed in Chapter 6.

Finally, the last research question dealt with user satisfaction. We have seen that participants rated the difficulty of test items on the usability questionnaire with a mean score of 2.65 on a 1-5 Likert-type scale. This coincides with what was observed in relation to user interactions with the interface, i.e. a relative difficulty of use experienced by the participants. Nevertheless, the answers provided on the post-test survey revealed a general positive attitude towards the DiCE. This suggests that despite the complexity and the usability problems presented by the current interface, most participants find it worth making the effort to learn to use it. In sum, the results seem to demonstrate that participants of the usability experiment welcomed and appreciated the usefulness of what constitutes a unique lexicographic product, an online combinatory dictionary of Spanish.

5.2.7 Summary: Usability of the DiCE interface

The first part of the present chapter described a usability study of the DiCE web interface, which aimed primarily at exploring how successfully potential users can interact with the different search options offered by the dictionary. Consequently, a detailed description of the DiCE interface was provided, followed by the presentation of the methodology applied in the study. Usability was assessed on the basis of three main parameters: effectiveness, efficiency and user satisfaction. Data related to these was obtained from measuring participants' performance on a usability test as well as tracking user actions in log files and collecting user opinions in a post-test survey.

Results showed that while the complexity of the interface caused certain difficulties, most participants adopted a favorable attitude towards DiCE and successfully carried out searches for a good number of test items. A detailed analysis of user actions allowed to identify specific usability problems, which can serve to define changes to be made in the design of the dictionary interface, as well as to reflect on the characteristics of electronic combinatory dictionaries and learning tools in general.

Clearly, in order to obtain a better picture of the usability of the DiCE interface, future experiments should be carried out where participants' reference skills are better controlled for, possibly with the inclusion of familiarization tasks. Finally, it should be emphasized that the methodology applied in the experiment implies that the test can be completed from the participants' home computers, which facilitates data collection considerably, and, therefore, it may be of interest for future user experiments.

5.3 Autonomous collocation error correction with a data-driven approach

The second half of the present chapter is dedicated to describe the results of an experimental study whose aim was to test to what extent SFL learners are able to make use of linguistic information presented in the form of concordance lines to correct different types of collocation errors. This research aim is especially relevant in the case of evaluating the potential efficacy of an online corpus tool targeting collocation learning. As it was discussed in Section 3.4.2.3, existing online learning resources created for foreign language learners provide combinatory information on the basis of corpus data. This implies that, although tools differ in their functions and in the specific ways they display information, in general, successful users need to have certain skills in interpreting authentic language data, as well as in contrasting it with their own production (see Milton and Cheng 2010, 34).

5.3.1 Aims of the study

Section 3.4.2.2 discussed the use of corpora in the L2 context, as a resource for teaching and learning collocations. It was noted that corpus data and concordances are often recommended as useful reference and learning tools (Higueras García 2006; Lewis 2000b; Moreno Jaén 2008; Woolard 2000), while there is certain concern as to language learners' skills and motivation when it comes to querying corpora or interpreting language data (Geluso 2013; Huang 2014; Moreno Jaén 2008; Yoon and Hirvela 2004). Furthermore, only a limited number of studies have attempted to explore what linguistic phenomena get noticed or learnt from concordance input, or what types of errors are effectively corrected through using concordance feedback. Consequently, as we have seen, evidence on autonomous collocation learning as well as studies exploring L2 learners' use of corpus data to verify combinations or correct collocation errors is rather scarce.

Here I will focus on the latter aspect, which is of crucial importance in estimating the potential efficacy of an automatic collocation learning tool aimed at improving language learners' written production through providing concordance feedback on collocation errors. More precisely, the goal of the study described here is to verify whether SFL learners are able to autonomously correct learner collocation errors encountered in the CEDEL2 corpus, and corresponding to different error types included in the typology presented in the previous chapter (see 4.3.2.1) with the help of concordance lines.

The specific aims pursued in the experimental study presented here are formulated in the following research questions:

1) Can SFL learners correct collocation errors autonomously with the help of concordance lines?

2) The correction of what error types poses more difficulty for the learners when presented with the concordance lines?
3) What concordance format is more suitable to provide feedback on collocation errors?
4) What problematic aspects of concordance feedback can be observed in learners' output?
5) How can concordance feedback be improved in order to better assist language learners in the revision of collocation errors?

5.3.2 Methodology

In the course of the present study participants had to complete a test in which they were asked to reformulate the highlighted segment of a sentence containing a collocation error, first, without any aid, and second, with the help of concordance lines. Subsequently, participants' answers were coded in order to carry out quantitative and qualitative analyses, and assess the viability of autonomous correction of collocation errors through the use of concordance feedback. The following subsections provide a description of the methodological aspects of the study, through introducing the questionnaire used, describing the participants and providing detailed information on how answers obtained in the test were analyzed.

5.3.2.1 The questionnaire and experiment set-up

The experiment described in this study adopted a methodology similar to that applied by Chambers and O'Sullivan (2004) and Wu, Witten and Franken (2010) (see 3.4.2.2.B) in that it involved the correction of marked errors. Participants were given a test containing a total number of 20 enhanced learner sentences, originally extracted from the annotated CEDEL2 corpus, with each sentence containing an erroneous collocation. The erroneous expressions included in the test were selected to constitute a representative sample of the different descriptive collocation error types identified in the learner corpus and included in the error typology presented in the previous chapter. Table 33 shows the different error types included in the test. The numbers in brackets represent the number of questionnaire items corresponding to each specific error type. For the original version of the questionnaire and the English translation of the instructions see Appendix C and Appendix D, respectively. Appendix E provides a summary of all collocation errors included in the test, together with the expected corrections.

Note that, as it was already discussed in 4.3.3.1 in relation to the error annotation process followed in the learner corpus study described in the previous chapter, a single erroneous collocation can contain more than one error instance, corresponding to different error types. This explains why the 20 erroneous combinations included in the error correction test corresponded to a total number of 22 error instances. For example, the incorrect combination **dimos bienvenidas* lit. 'we gave welcomes' (instead of *dimos la bienvenida* lit. 'we gave the welcome') contains two error instances. Firstly, the article of the base, obligatory for the correct formulation of the collocation, is missing, and, secondly, the base noun is in the plural form instead of the singular.

Error types		Sample questionnaire items and expected corrections
Lexical errors (13)	Incorrect collocate (6)	**capturar la atención* instead of e.g. *captar la atención* 'catch sb's attention'
	Incorrect base (2)	**cambiar la mente* 'change sb's mind' instead of e.g. *cambiar la idea/opinión* 'change sb's opinion'
	Synthesis (2)	**misinterpretaciones* 'misinterpretations' instead of e.g. *malas interpretaciones* 'wrong interpretations'
	Analysis (2)	**haciendo de cotilleo* lit. 'making gossip' instead of *cotillear* 'gossip'
	Collocation with an incorrect meaning (1)	*les da la gana* 'they feel like doing sg' (colloquial) instead of *tienen ganas de* 'they want to/wish to do sg'
Grammatical errors (9)	Number (2)	**dimos bienvenidas* lit. 'we gave welcomes' instead of *dimos la bienvenida* 'we gave a welcome'
	Gender (1)	**crímenes violentas* instead of *crímenes violentos* 'violent crimes'
	Governed preposition (2)	**montar una bicicleta* instead of *montar en una bicicleta* 'ride a bike'
	Determination (2)	**esperando un metro* instead of *esperando el metro* 'waiting for the metro'
	Pronoun (2)	**la película se trata de* instead of *la película trata de* 'the film is about'

Table 33 Collocation error types included in the test

As mentioned above, in the course of the experiment, participants were asked to revise the erroneous segments explicitly marked in the sentences twice. First they were instructed to try to reformulate the incorrect segments without any aid. This part of the experiment was considered to constitute a pretest, allowing to control for the cases when participants were able to provide correct answers when relying on their own knowledge. Second, participants were asked to correct the sentences again with the help of concordance data.

In order to test two different ways of presenting feedback, in the case of seven questionnaire items, half of the participants were provided with concordance lines in the form of full sentences (see (31)), while the other half were presented with corpus-derived n-grams (see (32)), similarly to the case of the FLAX interface (see 3.4.2.3), tested in Wu, Witten and Franken (2010). In the case of the remaining thirteen questionnaire items, all participants received full sentence concordances.

(31) La ingesta de líquidos es a lo que habitualmente **prestamos** menos **atención**.
Los niños sienten celos y los expresan a través de necesidades a las que los padres deben **prestar atención**.
Esta siempre ha sido la enfermedad que más me ha asombrado y que me ha **llamado** la **atención** de una forma especial.
Aquí os vamos a describir algunas opciones que nos han **llamado** la **atención** por sus cualidades.
Connelly **atrae** la **atención** y no deja que decaiga en ningún momento.
El otro día estaba comprándome libros en la librería y este me **atrajo** la **atención**.
Desarrollar una cierta habilidad para el coqueteo te puede ayudar a aprender a **captar** la **atención** e interés de las personas que te agradan.
Los niños suelen portarse mal para **captar** la **atención** de los padres.

(32) capten la atención de
captar la atención y la
para captar la atención de
concentrado la atención de los
concentra la atención de
de concentrar la atención de
capten la atención de los
concitan la atención de los
concitaban la atención de los
captan la atención de los

Full sentence concordance lines were retrieved from the *esTenTen* corpus (Renau and Kilgarriff 2013) available on the *Sketch Engine* interface (Kilgarriff et al. 2004). Example sentences were selected manually through the following procedure. Firstly, the *word sketch* for the base of the intended collocation(s) was obtained using the *Sketch Engine* (see 3.4.2.3). Secondly, the most frequent collocates expressing a meaning related to the one attempted to be conveyed by the erroneous combination were selected. Thirdly, concordance lines for the given word combinations were retrieved and ordered using GDEX, a feature of the *Sketch Engine* which allows filtering concordance lines in order to facilitate obtaining those that best exemplify an expression (Kilgarriff et al. 2008). Naturally, this strategy could not be used in the case of *analysis* errors, which consist of learners using a collocation-like expression (**haciendo de cotilleo* ‘making gossip’) instead of a single word L2 lexical item (*cotillear* ‘to gossip’). In these cases simple concordance searches were carried out to retrieve sentences illustrating the use of feasible target words, as well as lexical items with related meanings.

The corpus derived phrases or n-grams were obtained using a technology under development and aimed at offering automatic feedback on Spanish collocation errors (Ferraro et al. 2014). Given the elevated number of n-grams retrieved per error, each sample was manually filtered before its inclusion in the questionnaire.

Note that participants were not familiarized with the notion of collocation prior to taking the test. In addition, as opposed to some of the other studies dealing with the use of concordances in the classroom, where learners are introduced to DDL methodology as part of the experiment set-up (e.g. Geluso and Yamaguchi 2014; Moreno Jaén 2007), participants of this study did not receive any such instruction. Consequently they had to rely on their individual skills and previous learning experiences, much like when coming across a new learning tool and starting to explore its use in an autonomous manner.

5.3.2.2 Participants

A total number of 18 participants completed the questionnaire administered in the course of the study. They were all SFL students, who claimed to have intermediate to advanced (B1-C1) proficiency level of Spanish, see Figure 52. 12 participants were Erasmus students at the University of A Coruña, attending a Spanish language course in the Language Center of the same institution. Another 11 participants were students at the Escuela Oficial de Idiomas (EOI) in A Coruña. Five participants in the latter group, however, did not complete the entire questionnaire, therefore, their answers were excluded from the data analysis. Note that, in the case of the EOI students, the author was not granted direct access to the participants, therefore the questionnaires were administered by the language course teachers themselves.

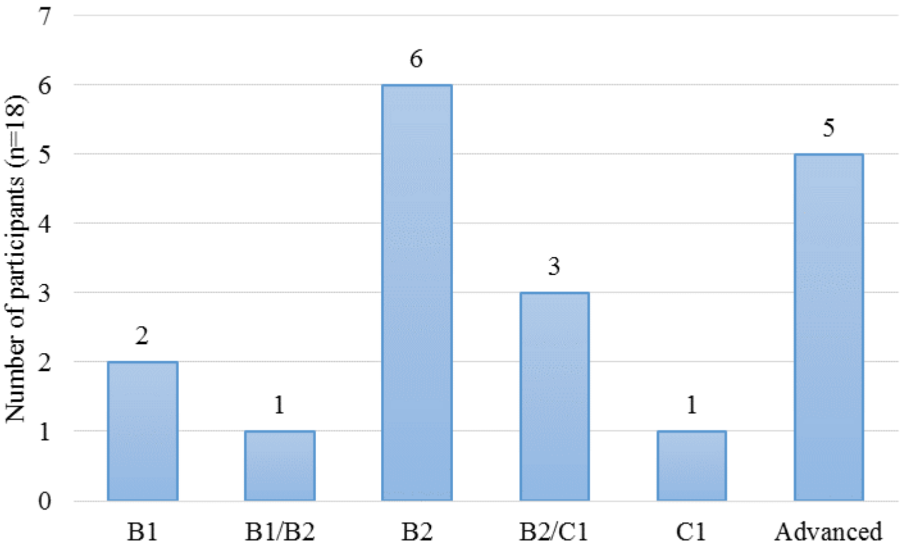


Figure 52 Participants' Spanish proficiency level according to self-assessment

The age of participants was between 20 and 40, with 15 out of the total number of 18 participants between the ages of 20 and 25. Most participants claimed to have been learning Spanish for various years, and all of them were residing in Spain at the time of the study. As for their native language, participants constituted a rather heterogeneous group, with ten different mother tongues, including Czech (1), English (3), French (1), Frisian/Dutch (1), German (4), Hungarian (1), Italian (2), Portuguese (3), Romanian (1) and Russian (1).

5.3.2.3 Data analysis

Participants' answers were coded according to a number of parameters in order to carry out a quantitative analysis. This subsection provides a brief explanation of these.

The first aspect taken into account was whether the answer provided by the participant constituted a suitable correction of the highlighted segment of the learner sentence included in the test, representing the erroneous collocation. Therefore, each answer provided in both iterations of the questionnaire was tagged as *correct* or *incorrect.* In order for a participant's answer to qualify as *correct*, it not only had to be comprised of the suitable lexical items and be grammatically correct, but it also had to fit the context to form a correct sentence, otherwise it was considered to be *incorrect*.

In the case of the second part of the test, in which participants had to revise the erroneous sentences with the help of corpus data, it was also taken into account whether the answer provided by a given participant differed from the one given on the same questionnaire item in the pretest. It was hypothesized that whenever the same answer was repeated by a participant in both iterations, it was not possible to decide whether the participant took the concordances into account when correcting the erroneous sentence in the second part of the test, or they assumed that they could provide a suitable answer relying on their own knowledge, and disregarded the corpus data all together. Consequently, as it is discussed when presenting the test results, *repeated answers* were discarded or considered separately in certain parts of the data analysis.

As it was mentioned above, participants' answers were only considered to be correct if the highlighted segments were not only reformulated to correct the collocation error, but they also fit the context. This implies that, the fact that a reformulation suggested by a participant was coded as incorrect, did not necessarily mean that the target collocation error, or at least one of the errors affecting the collocation was not corrected. For instance, in the case of test item 8 (shown in (33)), one participant substituted the erroneous segment for the correct collocation *dar las gracias* lit. 'to give thanks', but used an incorrect form of the verb, **demos* instead of *dimos* 'give-1pl-past', introducing a new error. In other cases, participants made changes to the original sentence that did not result in correcting the target error at all. For example, in the case of test item 9 (shown in (34)), one participant substituted the combination **futuro lejos* lit. 'far future' by another incorrect expression **futuro avanzado* lit. 'advanced future'. Therefore, it was deemed necessary to examine participants' answers in more detail regarding the changes made in an attempt to reformulate the highlighted segment.

(33) Al día siguiente, nos despedimos, y **gracias,** y caminamos hacia el puerto. 'The next day, we said good bye, and *thank you, and walked towards the port'.

(34) En cuanto al **futuro lejos**, también tengo muchas ideas. 'Regarding the *far future, I also have many ideas.

That is why participant answers were categorized regarding whether they constituted a *positive, negative* or *neutral* change with respect to the original sentence. Answers were labeled as containing *positive* change when at least one error affecting the collocation in the original sentence was corrected, and no new error was introduced. Answers where the participant's correction introduced a new error not present in the original sentence – as in the case of **future avanzado* – were considered to constitute a *negative* change, while cases where the participant corrected a collocation error affecting the collocation in the original sentence, but, at the same time, introduced a new error – as in the case of **demos las gracias* –, were coded as containing both *positive* and *negative* changes. Finally, answers that neither corrected any error affecting the collocation in the original sentence, nor introduced a new error were tagged as *neutral* changes.

Furthermore, each answer was tagged according to whether it contained the correction of the target collocation error found in the original sentence. This allowed to obtain quantitative data regarding the number of times a specific error type was corrected when concordance data was available. A collocation error was considered to have been likely corrected with the help of concordance lines only when participants had not provided a correction of the given error in the pretest, or when the concordances prompted them to provide a new correction, which was different from the one they had proposed relying on their own knowledge, such as e.g. the use of a different collocate. Consequently, in order to establish the number of cases when concordance data aided participants in the correction of target errors, in reality, three aspects had to be taken into account: 1) whether the answer provided with concordance feedback contained the correction of the error, 2) whether the answer was identical to the one given in the pretest, i.e. whether it was a *repeated answer* (see above), and, in the case of answers containing a reformulation in comparison to the pretest, 3) whether the error was corrected in the same way as in the case of the answer given in the pretest, i.e. whether it constituted a *repeated correction*, see 5.3.3.2.

Finally, in order to shed more light on the difficulties posed by the use of concordance feedback, new errors not present in the original sentences, i.e. introduced in participants' answers, were also tagged, taking into consideration whether they could be likely induced by the concordance lines.

5.3.3 Results and discussion

This section describes the results obtained from the analysis of language learners' responses on the questionnaire used to test their ability to correct collocation errors with the help of concordance feedback. The five research questions presented above will be considered one by one through relevant quantitative and qualitative data.

5.3.3.1 Overall effectiveness of concordance feedback

In order to assess the overall effectiveness of concordance feedback, the number of correct and incorrect answers in the pretest condition and the test condition was observed, as well as the changes participants made in the highlighted erroneous segments. As it is shown in Figure 53, when relying solely on their knowledge, participants managed to provide correct reformulations of the marked erroneous segments in the case of 33.61% of all test items, while they provided incorrect answers in 46.39% cases, leaving a total number of 72 (20%) questionnaire items unanswered. In comparison, when provided with corpus data, participants provided correct reformulations in 62.78% of cases, while the proportion of incorrect answers (31.39%) and questionnaire items left blank (5.83%) was lower than in the pretest. The Wilcoxon signed ranks test carried out with data referring to the amount of correct reformulations, i.e. test scores achieved by each of the participants in the two test conditions shows that the difference observed is statistically significant since the null-hypothesis can be rejected (W=6.5, z=-3.44, p=0.0006).

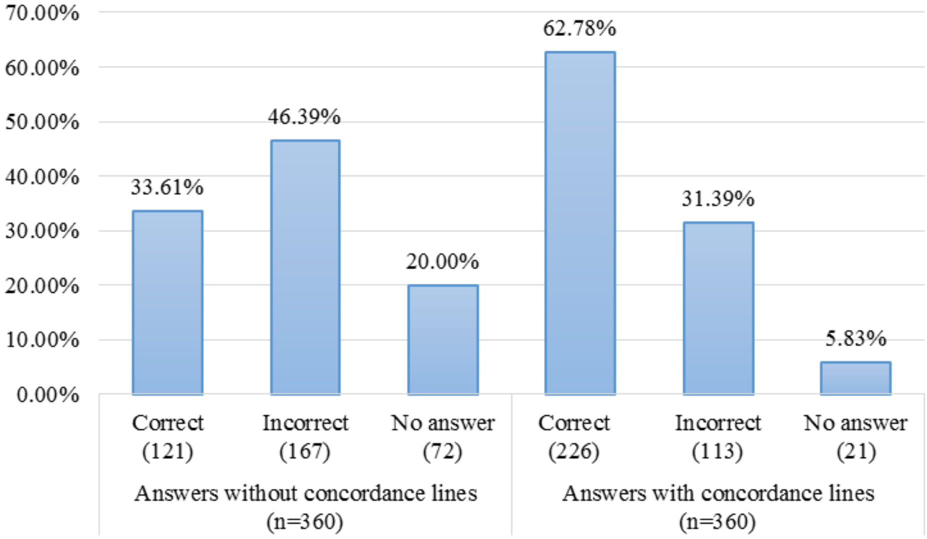


Figure 53 Number of correct and incorrect reformulations, as well as questionnaire items left blank under the pretest condition and the concordance feedback test

As it was suggested above, in order to better evaluate the effectiveness of concordances, cases when participants provided the same answer for a test item under the pretest and the concordance feedback conditions were also enumerated. When eliminating these repeated answers, constituting 24.72% of all answers, from the dataset, we are left with 153 novel correct answers (42.50%) as opposed to 94 (26.94%) incorrect reformulations, see Figure 54.

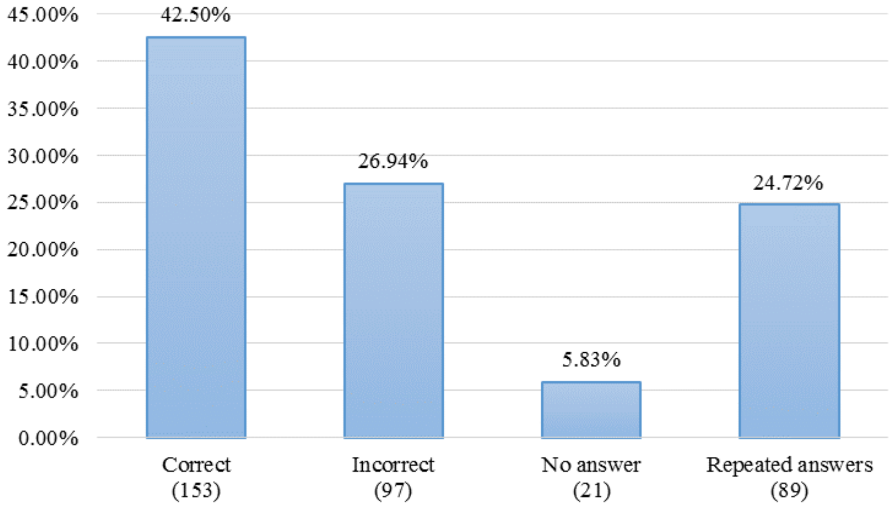


Figure 54 Number of correct and incorrect reformulations provided using concordance feedback when eliminating answers identical to those provided in the pretest

Figure 55 shows the percentage of positive, positive/negative, negative and neutral changes made by participants in each test condition. The numbers in brackets in the legend indicate the number of changes observed in the pretest and the concordance feedback condition respectively. From the total number of 289 reformulations provided in the pretest, 47.39% constituted a positive change with respect to the original sentence, through correcting at least one target error, while in 14.63% of the answers although at least one target error was corrected, a new error was also introduced (positive/negative change). 31.36% of reformulations introduced a new error without correcting the target collocation error and in 6.97% cases participant answers were judged to be neutral, as they did not correct the target error, nor did they introduce a new error in the highlighted segment. In comparison, out of the 250 answers provided when using concordance feedback, not constituting the repetition of reformulations offered in the pretest, target errors were corrected in a higher proportion. 67.20% of answers were classified as containing a positive change. In the case of 20.40% of answers a target error was corrected, while a new error was introduced. Negative and neutral changes were made in the case of a lower number, constituting 12% and 0.40% of reformulations respectively.

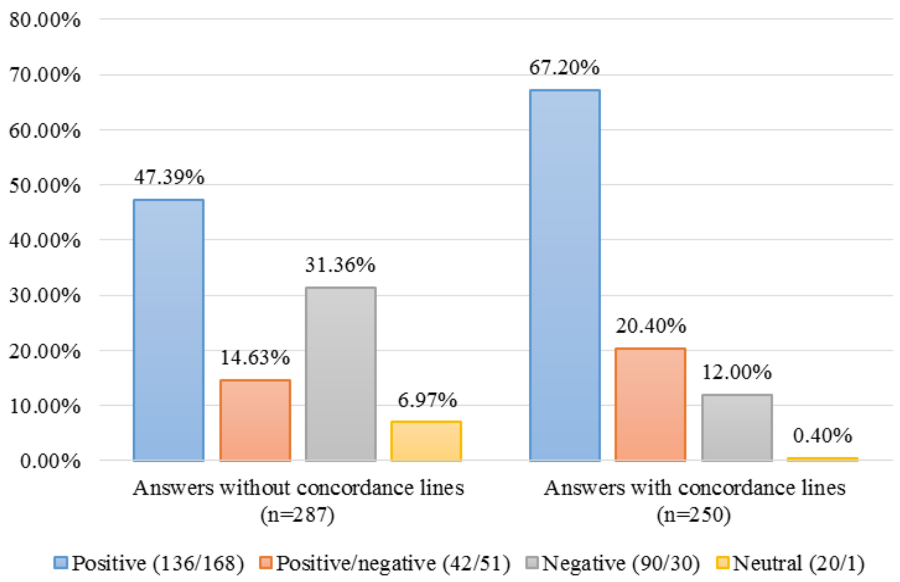


Figure 55 Amount of positive, positive/negative, negative and neutral changes in participants' reformulations

In sum, the data presented here suggests that concordance lines constitute an effective resource that learners can use autonomously for correcting errors, since participants were observed to produce a higher amount of entirely correct reformulations when provided with corpus data than without it, while the number of incorrect reformulations as well as test items left blank was lower. At the same time, the higher rate of positive and lower rate of negative and neutral changes made to the erroneous segment suggests that concordance lines constituted a useful tool for identifying errors in the original learner sentences. Nevertheless, the prevalence of negative changes in reformulations reveals that concordances did not help participants to overcome all aspects of production problems and/or the feedback itself might have induced the production of certain errors.

5.3.3.2 Effectiveness of concordance feedback in the case of different error types

Once it has been established that language learners are able to use concordance lines for autonomous error correction, we can turn to the second research question concerning the difficulty posed by the correction of different types of collocation errors. As it was noted above, the number of cases when concordance data likely aided participants in the correction of target errors was established taking into account three criteria. For an answer provided in the error correction test to be considered to contain a correct reformulation of the target error that likely resulted from concordance feedback, 1) it had to contain the correction of the target error, 2) it could not be identical to the answer given in the pretest, i.e. a *repeated answer*, and 3) it could not contain a *repeated correction*, i.e. the error could not be corrected in the same way as in the case of the answer given in the pretest.

Table 34 shows the percentage of errors corresponding to each error type that was successfully corrected or was not corrected in participants' answers, as well as cases when no answer was provided. In the case of the answers given in the test condition when concordance lines were provided, the percentage of repeated corrections and repeated errors is also indicated – note that, as explained above, the latter two cases should be excluded from the count when considering strictly the effect of concordance data on error correction. Numbers in brackets following the name of each error type indicate the number of corresponding error instances included in the test.

		Answers without concordance lines (n=360)			**Answers with concordance line (n=360)**				
		Corrected	Not corrected	No answer	Corrected	Not corrected	No answer	Repeated correction	Repeated answer
Lexical errors	Incorrect collocate (6)	44,44%	35,19%	20,37%	56,48%	10,19%	6,48%	1,85%	25,00%
	Incorrect base (2)	47,22%	36,11%	16,67%	52,78%	11,11%	5,56%	8,33%	22,22%
	Synthesis (2)	16,67%	58,33%	25,00%	72,22%	8,33%	2,78%	11,11%	5,56%
	Analysis (2)[40]	25,00%	44,44%	30,56%	58.34%	8,33%	11,11%	5,56%	16,67%
	Collocation with incorrect meaning (1)	27,78%	44,44%	27,78%	66,67%	5,56%	11,11%	0,00%	16,67%
	TOTAL (13)	**36,32%**	**41,03%**	**22,65%**	**59,40%**	**9,40%**	**6,84%**	**4,70%**	**19,66%**
Grammatical errors	Number (2)	36,11%	44,44%	19,44%	30,56%	25,00%	5,56%	22,22%	16,67%
	Gender (1)	50,00%	27,78%	22,22%	38,89%	22,22%	5,56%	0,00%	33,33%
	Governed preposition (2)	41,67%	52,78%	5,56%	38,89%	13,89%	0,00%	19,44%	27,78%
	Article (2)	38,89%	38,89%	22,22%	38,89%	16,67%	8,33%	16,67%	19,44%
	Pronoun (2)	61,11%	25,00%	13,89%	33,33%	2,78%	2,78%	11,11%	50,00%
	TOTAL (9)	**45,06%**	**38,89%**	**16,05%**	**35,80%**	**15,43%**	**4,32%**	**15,43%**	**29,01%**

Table 34 Participants' success in correcting specific collocation error types in the pretest and when presented with concordance lines

[40] In the case of one of the analysis errors included in the test, **me ponen muy apasionada* lit. 'they make me passionate' the concordance lines offered contained not only the single word item considered to be the most suitable correction, but also collocations. Three participants reformulated the erroneous segment using a suitable collocation (*sentir pasión* 'feel passion, enthusiasm', *despertar una pasión* 'wake a passion'), while six participants used the single word expression *apasionar* when working with the concordances.

As it is apparent from the data, while participants were in general more successful in correcting grammatical (45.06%) than lexical errors (36.32%) when relying on their own knowledge, they corrected a higher percentage of lexical errors (59.40%) successfully than grammatical errors (35.8%) when presented with the concordance lines, leaving aside repeated corrections and repeated answers. Accordingly, a higher amount of answers not containing a suitable correction of the target error were found in the case of grammatical errors (15.43%) than lexical errors (9.4%) when using concordance feedback.

As in the case of the overall test scores, the Wilcoxon signed-rank test was applied to establish whether the difference between the number of errors corrected by each participant in the pre-test and the concordance feedback conditions can be considered statistically significant. In order for this, I took into account all participant answers containing the correction of the target error produced using concordance feedback – regardless of whether they constituted repeated corrections. The results of the Wilcoxon signed-rank test showed that the difference in the performance of the participants in the two test conditions is significant both in the case of lexical (W=1, z=-3.57, p=0.0004) and grammatical errors (W=7, z=-3.29, p=0.0005). For a summary of corresponding data – including not only the number of corrected error instances, but also the number of participant answers not containing the correction of the target error and cases when no reformulation was proposed by the participant – is shown in Figure 56.

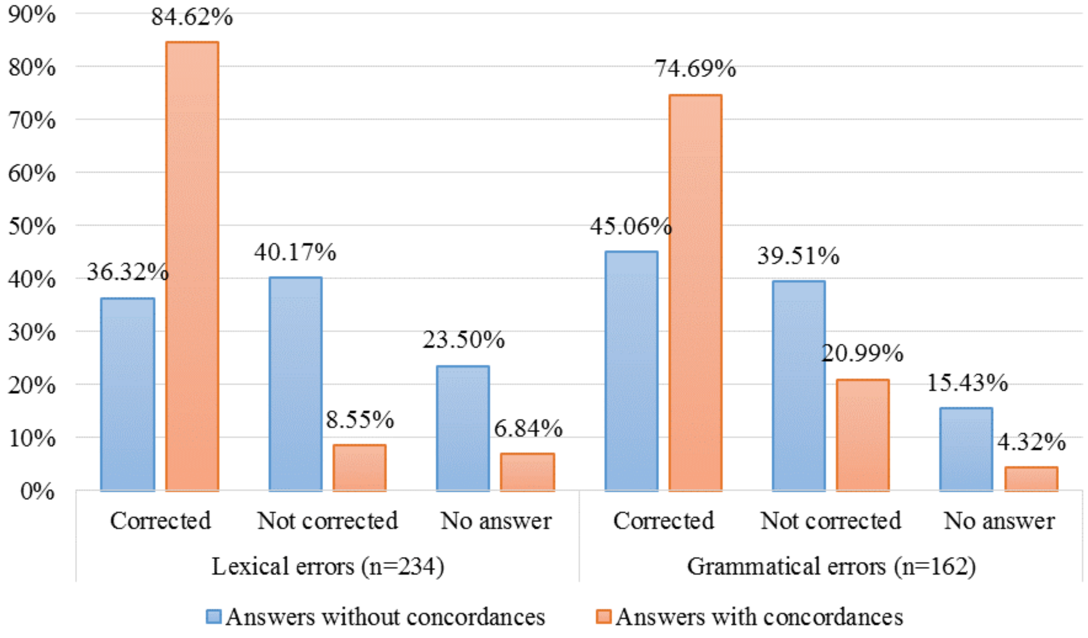


Figure 56 Amount of target lexical and grammatical errors according to whether a correction was suggested by participants in either of the two test conditions

Figure 57 provides a more visual summary of the success rate of error correction using corpus data – i.e. corresponding to the test condition when concordance lines were provided – observed in the case of individual error types. Note that, in contrast with Figure 56, percentages here are calculated through excluding repeated corrections and repeated answers, so that only reformulations most likely resulting from the consultation of the concordance lines are taken into account.

Again, it can be observed that participants were overall more successful in correcting lexical errors than grammatical errors. As for individual error types, it can be seen that participants appear to have been able to correct lexical errors with a rather uniform success rate, while individual types of grammatical errors differed considerably regarding whether they were successfully corrected. The lowest success rate was found in the case of number errors, with only 50% of answers generated through the likely use of concordances containing the correction of the target error, followed by gender (58.33%), and article errors (60.87%). The highest success rate was observed in the case of the correction of pronoun (85.71%) and governed preposition errors (73.68%). At the same time, note that, there was a relatively high percentage of answers (27.78%) given by participants when provided with concordance data, which did not contain the correction of the target error in the case of grammatical errors. This, together with the lower success rate in providing corrections shows that grammatical errors in general were much less salient than lexical errors, i.e. they were more difficult to identify probably due to the lack of explicit feedback concerning the nature of the error.

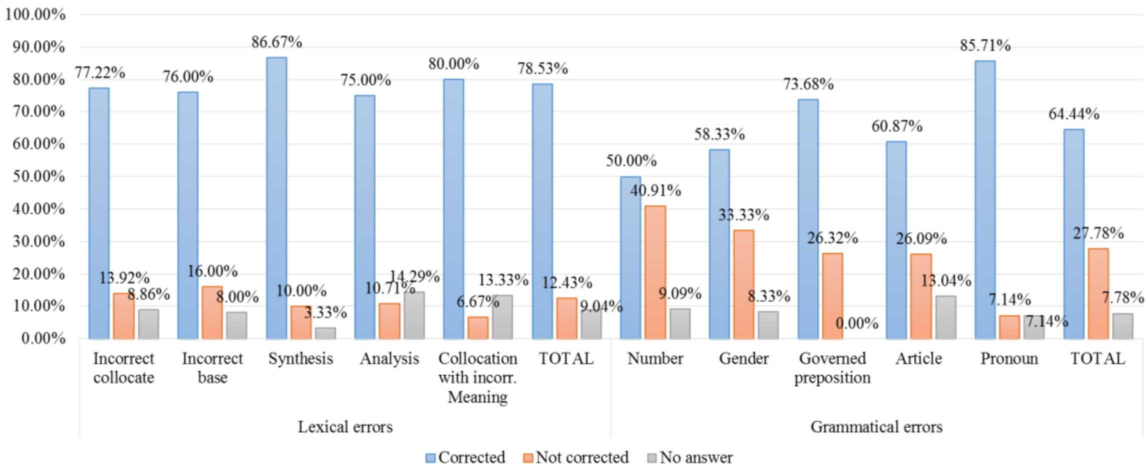


Figure 57 Participants' success in correcting different collocation error types when presented with concordance lines

The results of the present study do not coincide with the findings presented by Chambers and O’Sullivan (2004), who have found grammatical errors being corrected more often than lexico-grammatical errors including, for instance, incorrect verb+noun combinations, nor with Chambers and O’Sullivan (2006), where grammatical and lexico-grammatical errors were corrected at a similar rate. Note, however, that in these studies participants had to retrieve relevant concordances themselves through using a corpus tool, therefore, concordancing skills, as well as the fact that correct alternatives are in general more difficult to query in the case of certain lexical errors might have affected the results. On the contrary, the findings presented here are comparable to the results obtained by Wu, Witten and Franken (2010), who, in fact, focused on a similar set of error types as the study described here. The authors found that in the case of errors for which suitable feedback could be obtained using the *FLAX* learning tool – as confirmed by queries carried out by the researchers –, the success rate of error correction was 67%, with a higher rate of successful corrections in the case of lexical (70.5%) than grammatical errors (55%).

5.3.3.3 Effectiveness of different concordance feedback formats

The third research question the present study aimed to deal with concerned the effectiveness of different concordance formats in collocation error correction. As it was explained in 5.3.2.1, in the case of seven out of the 20 questionnaire items (see items 4, 5, 6, 7, 11, 12 and 15 in Appendix D), half of the participants were provided with concordance lines in the form of full sentences, while the other half were presented with corpus derived phrases or n-grams.

Figure 58 shows the overall number of correct and incorrect reformulations as well as questionnaire items left blank in the case of each feedback format as compared to the answers provided by corresponding participants in the pretest. While the two participant groups seem to have performed similarly in the pretest, when working with corpus data, their performance seems to differ. Namely, in the case of the group of participants who were given full sentence concordances, the number of correct reformulations increased to a greater extent as compared to the pretest (from 26 to 42) than in the case of participants working with the n-gram format (from 27 to 31). At the same time, participants using n-gram format concordances provided a higher number of incorrect answers (28) than the full sentence concordance group (19). Note however, that as shown by the results of the

Mann-Whitney U test the null hypothesis cannot be rejected neither in the case of the number of correct (Mann-Whitney U= 26.6, n1=n2=9, p>0.05 two-tailed) nor in the case of erroneous reformulations (Mann-Whitney U= 25.5, n1=n2=9, p>0.05 two-tailed). Therefore, there is not enough evidence to claim that the differences observed between the two participant groups receiving feedback corresponding to different concordance formats are statistically significant.

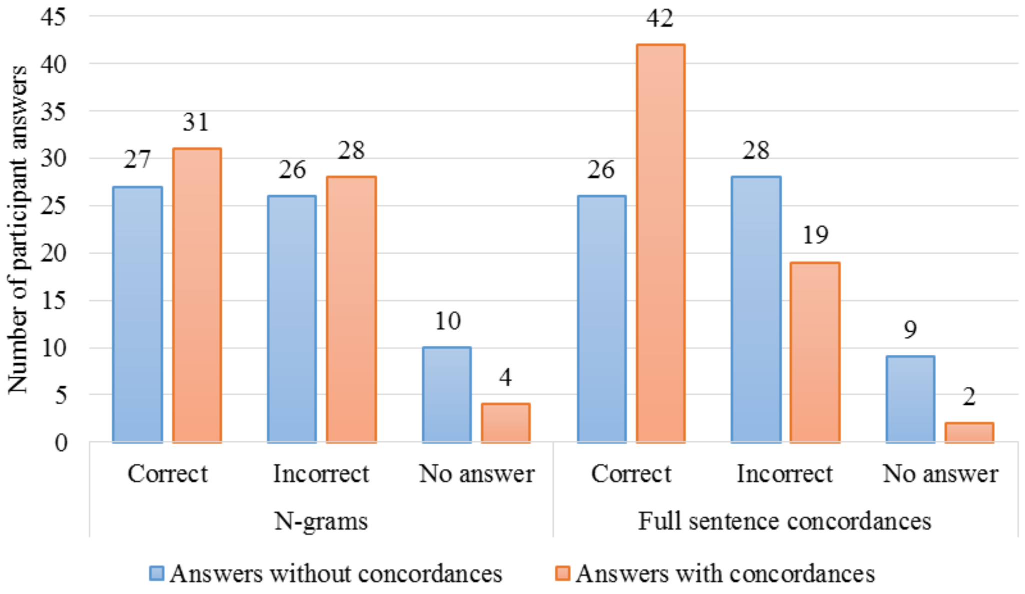


Figure 58 Number of correct and incorrect reformulations as well as items left blank according to concordance-format

Figure 59 shows the distribution of participant answers through eliminating repeated answers in the concordance condition, with the aim of focusing only on novel reformulations, which likely resulted from participants' use of the concordance feedback. It can be observed that while the amount of novel correct answers – as compared to the ones provided in the pretest condition – is similar in the case of the two groups, the number of incorrect reformulations is higher in the case of the group that received concordances in n-gram format (27 answers) than the one that was provided with full sentence concordances (15 answers).

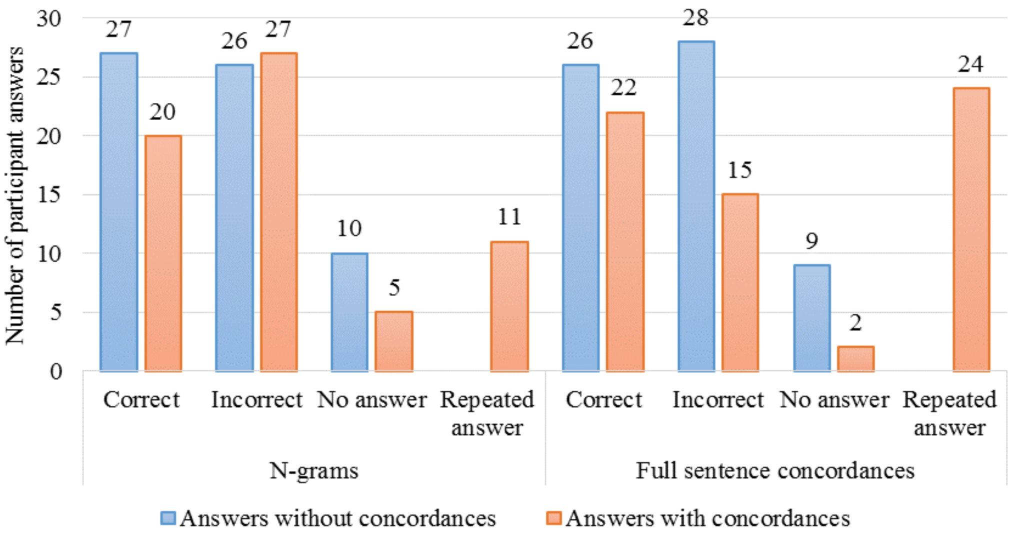


Figure 59 Number of correct, incorrect reformulations, items left in blank and repeated answers according to feedback type

Figure 60, representing the nature of changes introduced in participants' answers allows for similar conclusions since, while the number of answers containing positive changes is similar when working with n-grams and full sentence concordances, the number of answers containing negative and positive/negative changes is slightly higher in the first case. Notably, the number of answers containing negative changes decreased more drastically as compared to the pretest in the case of participants who were provided with full sentence concordance lines.

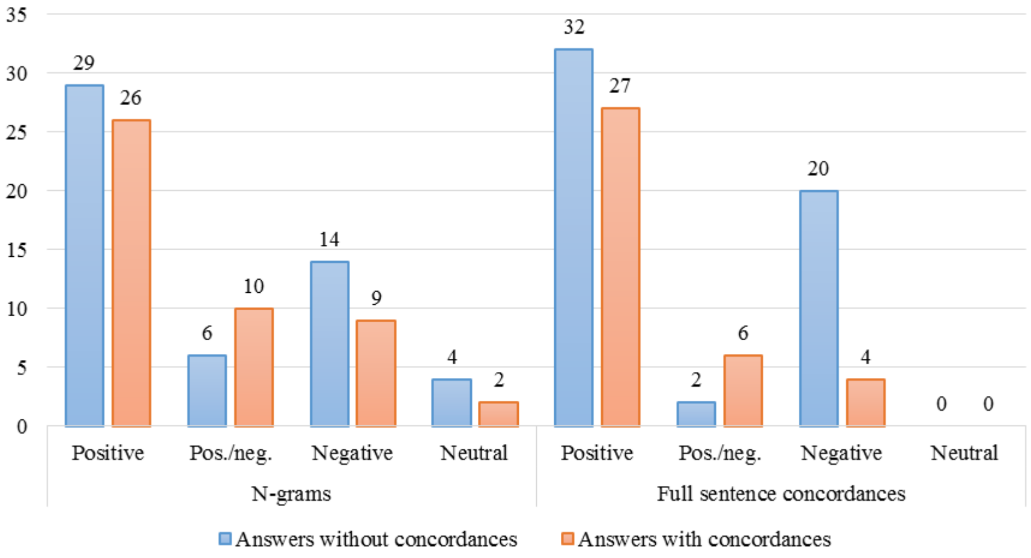


Figure 60 Number of answers containing positive, positive/negative, negative and neutral changes according to feedback type

5.3.3.4 Problematic aspects of concordance feedback

In order to explore potential problems arising from the use of concordance feedback, constituting the target of the forth research question, I carried out a qualitative analysis focusing on incorrect answers provided in the error correction test. As explained in 5.3.2.3, participant answers were coded as incorrect when they introduced new errors not present in the original sentence, or constituted incomplete corrections of the original learner utterance presented in the test. It was observed that these erroneous answers can be grouped in four main categories concerning the relationship between the string produced by the participant and the concordance lines themselves.

Firstly, in a few cases, participants did not seem to have made use of the concordance lines, since their answers consisted of segments which were not exemplified in the concordances, see Table 35. This suggests that, when comparing the erroneous learner sentence constituting the test item to the concordance lines provided, participants did not manage to identify the error and/or its suitable correction. Since the new erroneous forms did not seem to be related to the content of the concordances, they were described as non-concordance induced errors.

Erroneous segment in the original learner sentence	**Incorrect answer provided by participant**	**Expected correction**
*misenterpretaciones ‘misinterpretations’	*mis interpretaciones ‘my interpretations’	malas interpretaciones ‘misinterpretations’
nos despedimos, y *gracias, y caminamos hacia el puerto lit. ‘we said goodbye, thanks, and we started to walk towards the port’	nos despedimos, y *gracias a Dios caminamos hacia el puerto ‘we said goodbye, and thank God we started to walk towards the port’	nos despedimos, y *les* dimos gracias, y caminamos hacia el puerto ‘we said goodbye, and thanked them, and started to walk towards the port’
*dimos bienvenidas a los nuevos estudiantes, y dijimos adiós a los que se iban lit. ‘we gave welcomes to the new students, and said goodbye to the ones who were leaving’	*hemos dado bienvenidas a los nuevos estudiantes, y dijimos adiós a los que se iban lit. ‘we have given welcomes to the new students, and said goodbye to the ones who were leaving’	dimos la bienvenida a los nuevos estudiantes, y dijimos adiós a los que se iban lit. ‘we gave the welcome to [we welcomed] the new students and said goodbye to the ones who were leaving’

Table 35 Non-concordance induced errors in participants’ answers

In the case of the first example shown in Table 35, participants were expected to correct the erroneous form **misenterpretaciones*, likely constituting a borrowing of the English word *misinterpretation* (see 4.3.2.1.B). Although the concordance lines provided in the test contained the collocations *interpretación errónea* ‘erroneous interpretation’, *interpretación incorrecta* ‘incorrect interpretation’ and *mala interpretación* ‘false interpretation' (see Appendix D), all of which constituted a suitable correction of the error, the participant in question seems to have interpreted the erroneous segment as containing a spelling error, not realizing that the attempted meaning should be correctly expressed by a multiword expression.

In the case of the second group of erroneous answers, participants seemed to have identified the target error contained in the test item, however they did not manage to pick the suitable option to correct it on the basis of the examples provided in the concordance lines. In these cases participants used expressions or word combinations extracted from the concordances conveying a meaning which (slightly) differed from the one required in the original sentence (see Table 36). This type of erroneous reformulations might have resulted from the lack of sufficient context either in the sentence constituting the test item, or in the concordances themselves. Since the expressions incorrectly used in participants' answers came from the concordance lines, these cases can be described as meaning related concordance induced errors.

In the case of the last example shown in Table 36, the governed preposition *en* was missing from the collocation in the original learner sentence, while a superfluous indefinite article was inserted (**montar* [*en*] *una* *bicicleta*). As it is shown, one of the participants opted for substituting the target collocation by the combination *utilizar una bicicleta* 'use a bike' illustrated in the concordance lines, which is not suitable in the given context.

Erroneous segment in original learner sentence	**Incorrect answer provided by participant**	**Expected correction**
En cuanto al *futuro lejos, también tengo muchas ideas. lit. 'As for the far future, I also have many ideas.'	En cuanto al *futuro cercano, también tengo muchas ideas. 'As for the near future, I also have many ideas.'	En cuanto al futuro lejano, también tengo muchas ideas. 'As for the distant future, I also have many ideas.'
Los *derechos mujeriles empezaban a mejorar en casi todos los regiones del mundo… lit. 'Womanly rights started to improve in almost every region of the world…'	Los *derechos humanos empezaban a mejorar en casi todos los regiones del mundo… 'Human rights started to improve in almost every region of the world…'	Los derechos de la mujer empezaban a mejorar en casi todos los regiones del mundo… 'Women rights started to improve in almost every region of the world…'
Mi futuro no*tiene limitades. 'My future has no limits.'	Mi futuro no*tiene limitaciones. lit. 'My future has no limitations.'	Mi futuro no tiene límites. lit. 'My future has no limits.'
Me gustaría *montar una bicicleta en el bosque tropical. 'I would like to ride a bike in the rain forest.'	Me gustaría *utilizar una bicicleta en el bosque tropical. lit. 'I would like to use a bike in the rain forest.'	Me gustaría montar en bicicleta en el bosque tropical. 'I would like to ride a bike in the rain forest.'

Table 36 Meaning related concordance induced errors

The third category of erroneous segments includes cases where participants inappropriately applied a pattern derived from the concordance lines. In some of these cases participants failed to identify an error in the test item and, instead, adopted an expression or pattern from the concordances introducing a new error in their reformulation.

Erroneous segment in original learner sentence	**Incorrect answer provided by participant**	**Expected correction**
Basta con estar en una parada *esperando un metro para ver a los fumadores disfrutando de un cigarrillo… It is enough to be at a station waiting for a metro to see smokers enjoying a cigarette..	Basta con estar en una parada *mientras esperando un metro para ver a los fumadores disfrutando de un cigarrillo… It is enough to be at a station *while waiting for a metro to see smokers enjoying a cigarette… • *Error introduced*: adverb • *Source*: Mientras esperan el metro	Basta con estar en una parada esperando el metro para ver a los fumadores disfrutando de un cigarrillo… It is enough to be at a station waiting for the metro to see smokers enjoying a cigarette…
Cuando alguien dice que los gays no deben *tener los derechos para casarse… lit. 'When somebody says that gays shouldn't have the rights to get married…'	Cuando alguien dice que los gays no deben *tener los mismos derechos de non-gays… lit. 'When somebody says that gays shouldn't have the same rights as non-gays…' • *Error introduced*: missing argument (de casarse 'to get married') • *Source*: tener los mismos derechos de	Cuando alguien dice que los gays no deben tener derecho a casarse … 'When somebody says that gays shouldn't have the right to get married…'
*La película se trata de una mujer soltera, su hija y sus amigas… 'The film is about a single woman, her daughter and her friends..'	*La película, que se trata de una mujer solera, su hija y sus amigas.. 'The film, which is about a single woman, her daughter and her friends..' • *Error introduced*: subordinate clause • *Source*: películas que tratan de la	La película trata de una mujer solera, su hija y sus amigas.. 'The film is about a single woman, her daughter and her friends..'
*dimos bienvenidas a los nuevos estudiantes, y dijimos adiós a los que se iban lit. 'we gave welcomes to the new students, and said goodbye to the ones who were leaving'	dieron la bienvenida a los nuevos estudiantes… lit. 'they gave the welcome to [welcomed] the new students… • *Error introduced*: person agreement • *Source*: dieron la bienvenida en el	dimos la bienvenida a los nuevos estudiantes, y dijimos adiós a los que se iban lit. 'we gave the welcome to [welcomed] the new students…

Table 37 Concordance-induced errors involving the inappropriate application of a pattern

For instance, in the first example shown in Table 37, the participant introduced the adverb *mientras* 'while' erroneously used in the given context, while failing to notice the target error being the use of the indefinite instead of the definite article. In other cases, although the target error was corrected in the answer provided by the participant, a new error was introduced as a result of the straightforward reproduction of a pattern found in one of the concordance lines. For example, in the case of the last example shown in Table 37, both target errors found in the highlighted segment – the use of the plural, instead of the singular form of the noun *bienvenida* 'welcome' and the lack of the definite article – were corrected in the reformulation provided by the participant, nevertheless, the participant also reproduced the verb form found in the concordance line, *dieron* 'gave-3pl', which constitutes an error, given that the context required the first person plural form. It was also noted that novel errors resulting from adopting inappropriate patterns from concordances were observed more frequently when participants were provided with n-gram type concordance feedback.

Finally, in the case of the last category of incorrect answers established, participants managed to identify and correct the target error or one of the target errors found in the test item, however the suggested reformulation was not entirely correct, since they failed to apply fully the pattern exemplified by the concordance lines.

Erroneous segment in original learner sentence	**Incorrect answer provided by participant**	**Expected correction**
*dimos bienvenidas a los nuevos estudiantes, y dijimos adiós a los que se iban lit. 'we gave welcomes to the new students, and said goodbye to the ones who were leaving'	*dimos las bienvenidas a los nuevos estudiantes lit. 'we gave the welcomes to the new students *Remaining error*: plural *Source*: La ciudad dará la bienvenida a los estudiantes	dimos la bienvenida a los nuevos estudiantes lit. 'we gave the welcome to the new students
Cuando alguien dice que los gays no deben *tener los derechos para casarse… lit. 'When somebody says that gays shouldn't have the rights to get married…'	Cuando alguien dice que los gays no deben *tener derecho para casarse lit. 'have right to' *Remaining error*: governed preposition *Source*: El cónyuge y los hijos a cargo tienen derecho a ejercer una actividad económica	Cuando alguien dice que los gays no deben tener derecho a casarse.. lit. 'have right to'
…y entonces *encendió el fuego que quemó la casa y los mató. lit. '… and then she lit the fire that burnt the house and killed them.'	…y entonces prendió el fuego que quemó la casa lit. 'and then she set the fire that burnt the house *Remaining error*: prepositional complement *Source*: algunos manifestantes habían prendido fuego a un edificio	…y entonces prendió fuego a la casa… lit. '… and then she set fire to the house'

Table 38 Incomplete correction of highlighted segments in learner sentences

In the case of the second example shown in Table 38, the participant adopted the usage pattern of the collocation exemplified in the concordance lines – one of which is shown here – only partially, in that the noun *derecho* 'right' is being used in the singular and without the definite article, but the governed preposition used is still incorrect. A similar case can be observed in the last example shown in Table 38, where, in order to reformulate the original learner utterance correctly it was not enough to substitute the collocate verb *encender* 'light a fire' by *prender* 'set fire to', but the sentence structure also had to be adapted to the new verb. As it can be seen, the original erroneous sentence contained a relative clause (*que quemó la casa* 'that burnt the house'), nevertheless, as exemplified by the concordance lines, the verb *prender* requires a prepositional complement (*prender fuego a la casa* 'set fire to the house'), which the given participant failed to introduce.

5.3.3.5 Possible ways to enhance concordance feedback

The fifth and final research question deals with potential ways to improve concordance feedback. This section reviews he results of the test described above allow in order to draw a number of conclusions in this regard.

To begin with, as we have seen, participants were in general less successful at correcting grammatical errors than lexical errors with the help of concordance feedback. This can be most likely put down to the fact that grammatical errors are less salient when it comes to contrasting learners' own production, or in the case of the test carried out here, erroneous sentences produced by other language learners, with the corpus examples serving as a linguistic model. One way to make grammatical patterns visually more salient is to organize concordance feedback in a way similar to what has been proposed by Wu, Witten and Franken (2010). In the tool developed by these authors the occurrences of each collocation are grouped according to the pattern they represent, and are presented as short phrases or n-grams, together with information on the frequency of occurrence of each pattern (see 3.4.2.3.A). Similar solutions are offered in what were described as *pattern-search tools* in 3.4.2.3.B.

However, it should also be noted that, as it was discussed above, in the study described here participants presented with corpus data in n-gram format performed somewhat worse at correcting collocation errors than those provided full sentence concordances. As suggested, one likely reason for this is the lack of context in the case of these shorter phrases. In fact, we have seen that one type of concordance-induced errors found in participants' answers were precisely attributed to this cause. In addition, in previous studies investigating learners' attitudes, lack of wider context was claimed to be one of the difficulties when it comes to interpreting concordance lines (e.g. Huang 2014, 74). Consequently, it is desirable in all cases to allow access to full sentence concordances, or even broader contexts. At any rate, clearly, the ideal format for presenting concordance feedback in order to better promote the noticing of different linguistic features should be further investigated.

Besides limited context, another feature of concordance feedback which was described as contributing to incorrect reformulation attempts is the lack of explicit feedback concerning the exact characteristics of the error. Participants in the study had to actively compare the test item with the concordances serving as a model, notice relevant differences, deduce the suitable pattern, or choose the lexical items to be used in the given context, and adapt the linguistic elements to the sentence to be corrected. Precisely this active role of the language learner is what has been seen as one of the key beneficial aspects of using authentic language data and promoting inductive learning (see e.g. Johns 2000; Yoon and Hirvela 2004). Nevertheless, as it was suggested by the results presented here, language learners often have difficulty in identifying an error on their own account. A way to provide learners with the opportunity to explore language data for themselves, as well as with more explicit pointers concerning the correct target forms, would be to offer multi-step feedback. A tool devised as a writing tutor could as a first step provide relevant language data, offering the user the opportunity to contrast it with their own production autonomously, while, as a second step, it could provide more explicit, additional aid.

Finally, a further problematic aspect of concordance line feedback is that corpus data does not necessarily contain all the information necessary for the correction of a given error. Certain aspects may not be sufficiently exemplified, or may not be easily deduced from the data, such as e.g. irregular verb forms, the use of governed prepositions. In addition, learners may have problems interpreting authentic language due to a higher amount of unknown vocabulary. Therefore, it might be convenient to integrate a writing tool relying on corpus data with other resources, such as dictionaries.

5.3.4 Summary: Correction of collocation errors with concordance data

The second part of this chapter described a study aimed at testing language learners' ability to make use of authentic language data presented in the form of full sentence concordances or n-grams to correct different types of collocation errors. In order for this, sentences containing erroneous combinations were retrieved from the CEDEL2 corpus to constitute a sample which was deemed to be representative of the types of descriptive collocation errors identified previously through studying SFL learners' collocation production (see Chapter 4). The feasibility of SFL learners correcting these errors relying on concordance lines was assessed in a pen and paper test, the results of which were analyzed to establish the success rate in error correction, as well as to explore potential difficulties posed by concordance feedback.

Results showed that SFL learners were in general able to make use of language data presented in the form of concordances to correct the target collocation errors, while it was observed that participants were more successful in correcting lexical errors than grammatical errors. A tentative comparison of the effectiveness of n-gram and full sentence concordance formats showed that the latter resulted in more successful error correction attempts. Erroneous answers provided by participants were considered to provide an insight into the problematic aspects of interpreting concordance data, related to limited context and lack of explicit indication of the error. Finally, possible ways to enhance concordance feedback were discussed taking into account the results obtained in the experiment.

Note that the results of this study are promising to the extent that evidence is provided as to SFL learners' ability to autonomously interpret authentic language data and contrast their production with it. However, since the experiment was carried out using concordance handouts, it does not provide information regarding the attitudes or the motivation of learners in using actual concordance software or a learning tool more specifically tailored to their needs. In this regard, Gaskell and Cobb (2004), for instance, found that learners were rather reluctant to use concordance tools when they were not explicitly instructed to do so. Another aspect subsequent work should attend to is a more rigorous comparison of different formats of presenting corpus data, while it would also be of interest to find evidence concerning the long term effects of the autonomous error correction on language learning.

5.4 Summary

This chapter described two experimental studies, both aiming to examine how successfully learners can interact with resources that can be potentially used as reference tools when producing collocations: DiCE, an online collocation dictionary of Spanish, and concordance lines representing authentic language data. As it was mentioned, in comparison to learners of English, SFL learners have less learning resources at their disposal, while empirical research into the use and effectiveness of learning tools available in Spanish is also more scarce. At the same time, this type of research is considered to be useful when it comes to designing learning resources, or aiming to enhance the features of existing ones.

The results of both the usability experiment aimed to test the online DiCE interface and the error correction test making use of concordance data demonstrate that language learners are in general able to use both types of tools successfully. Although, it cannot be objectively determined to what extent observed success rates are satisfactory, in the case of DiCE, we have seen that participants' answers on the post-test survey expressed a mainly positive attitude towards the tool.

Importantly both studies provide an insight into how learners interact autonomously, i.e. without any external aid, with potential learning tools. Furthermore, the results can be interpreted as pointers highlighting problem areas where there is room for improvement concerning the design of these. The different ideas concerning the enhancement the online collocation dictionary and the design of a learning tool offering concordance feedback discussed here are further developed in the following chapter.

Chapter 6. Towards a new collocation learning tool

6.1 Introduction

The previous chapters of this thesis provided an overview of different resources which can be exploited for collocation learning (see 3.4.2), described an empirical study of SFL learners' collocation production (Chapter 4), as well as two experimental studies concerned with learners' use of learning resources (Chapter 5). The goal of this chapter is to present a proposal of a new online collocation learning tool aimed at learners of Spanish and based on the findings discussed until this point.

The online tool presented here is designed to constitute an interactive learning environment. Its primary function is that of a reference tool which helps learners of Spanish resolve production problems related to word combinations. At the same time, it aims at providing a personalized learning experience, through offering users the option of creating a personal collocation notebook or collocation dictionary, in which learners can collect the word combinations they have a special interest in, as well as through allowing users to generate collocation activities focusing on the combinations they wish to practice or commit to memory. The following sections are dedicated to discussing the different functions of the proposed learning tool. Note that the content of this chapter is limited to describing the tool strictly from the point of view of design and usability, consequently, the discussion of technical details is kept to the minimum.

6.2 A comprehensive reference tool for collocation learning

The principal function of the collocation learning resource proposed in this chapter is to serve as a comprehensive reference tool. Such tool is envisioned through the integration of data coming from a collocation dictionary, more specifically, DiCE (Alonso Ramos 2004), and a corpus-based collocation search tool, such as *HARenEs* (Alonso Ramos, García Salido, and Vincze 2014; Wanner, Verlinde, and Alonso Ramos 2013; Wanner et al. 2013). As we have seen in 3.4.2.1, the number of lexicographical works dealing with collocations available to learners of Spanish is scarce, with the only freely available online dictionary being precisely DiCE, whose content is limited to the semantic field of emotion nouns. At the same time, section 3.4.2.3 presented an overview of tools which render corpus data accessible to language learners through extracting combinatorial information. The only such tools available from Spanish I am aware of are the *Sketch Engine* (Kilgarriff et al. 2004; Kilgarriff et al. 2014), the automatic Spanish collocation dictionary generated using the *word sketch* technology (Kilgarriff et al. 2008), and *HARenEs*.

Taking into consideration the fact that the elaboration of a comprehensive collocation dictionary requires considerable time and effort, technologies relying on corpus data are especially useful. The potentially large amount of combinations obtained from a large language corpus can be presented in a structured format and used to complement existing dictionary content (see e.g. Wanner 2006). This is in line with current trends in lexicography, where dictionary and language corpora are becoming more and more integrated, in the sense that corpora are not only consulted by the lexicographer to obtain linguistic data and examples to be included in the dictionary, but, in some cases, they are incorporated in the dictionary, offering additional information to the dictionary user (see e.g. Alonso Ramos 2009; Asmussen 2013; de Schryver 2003, 167–169). Consequently, it can be claimed that the incorporation of information extracted from corpus in the reference tool proposed here constitutes an efficient way of both extending and enriching the combinatorial information currently offered in DiCE.

As it is explained below in more detail, information coming from the collocation dictionary and the corpus tool are proposed to be treated as separate entities, such that while the two resources are searched simultaneously, results coming from each are displayed in parallel on a shared screen. In what follows, I first discuss the data categories and the types of information to be represented in what can be termed as the *Collocation dictionary module* and the *Collocation corpus module* respectively. Following this, I describe the query interface, as well as the types of feedback offered in the case of each potential search scenario, exemplified by different types of potential user queries.

6.2.1 Collocation dictionary module

Section 3.4.2.1.B and 3.4.2.1.C provided a detailed account of the way the content of collocation dictionaries is organized, and how collocations are described in these dictionaries. My aim here is to reconsider the types of information involved in describing word combinations. Firstly, I illustrate how an electronic collocation dictionary of Spanish can provide a more flexible access to combinatory information through regrouping combinations using semantic classes. Following this, I discuss in detail the necessity of a sufficiently detailed description of collocations, which may help learners to choose more native-like combinations, and use them more accurately.

6.2.1.1 Semantic classification of collocations in the dictionary

As we have seen in 3.4.2.1.B, collocation dictionaries generally list collocations in the entry of the base, and organize them in groups according to their syntactic pattern. In most dictionaries, combinations are further grouped according to their meaning, which in some cases is specified explicitly. It was also discussed in 3.4.2.1.F that this type of organization of dictionary entries is suited to a specific type of look-up situation, when the dictionary user has in mind both a given syntactic pattern to be used and a meaning to be conveyed. This type of look-up can be exemplified by the following question: 'What adjective(s) can I use to speak about the growing intensity of *anger*?'

It was also mentioned that an alternative approach, proposed in Jousse (2010), makes use of Lexical Functions (LF, Mel'čuk 1996), introduced within the framework of the Explanatory and Combinatorial Lexicology (Mel'čuk, Clas, and Polguère 1995) to create a comprehensive systematic semantic taxonomy applicable for the classification of collocations in the dictionary. This allows to group collocations according to their meaning, independently of their syntactic pattern, which implies that, once the semantic taxonomy is implemented in, for instance an electronic dictionary interface, a new access route to combinatorial information can be created. This access route was illustrated with the question 'What expressions can I use to speak about an increase of the intensity of *anger*?'.

Clearly, through establishing independent semantic and syntactic typologies for collocations, it is possible to render dictionary access more flexible. In addition, the option of querying collocations through their meaning may be beneficial for the learning process. Firstly, it allows learners to discover the full range of expressions corresponding to a given meaning, including those which do not match the syntactic pattern of L1 equivalents. Secondly, encouraging the learner to access collocations through their generic meaning, instead of the part of speech of their components, may help establish a sense of restricted lexical selection, emphasizing that the given combinations are not constituted by two fully autonomous lexical elements. Consequently, semantic access may contribute to overcoming the heavy reliance on L1 combinations, often resulting in erroneous collocations characteristic of learners' language production.

With the aim of demonstrating the potential of a semantically motivated grouping of collocations, I identified seven main semantic classes, some containing further subdivisions, which are fairly productive in the case of the VERB+NOUN$_{\text{COMP}}$, NOUN$_{\text{SUBJ}}$+VERB and NOUN+ADJECTIVE collocations containing emotion nouns included in DiCE. Note that these classes do not cover all the combinations included in the dictionary corresponding to the given patterns. The semantic classes were established on the basis of the LFs and semantic glosses used to describe the meanings of combinations in the dictionary (for a summary of the semantic classes see Table 39).

The first class referring to INTENSITY includes four subclasses which group collocations according to whether the emotion represented by the base of the combination is described as having HIGH INTENSITY (e.g.*matar a alguien de aburrimiento lit.* 'kill sb of boredom'), or LOW INTENSITY (e.g. *escasa consideración* 'little consideration'), or whether its intensity is GROWING (e.g. *fortalecer una amistad* 'strengthen a friendship'), or DECREASING (e.g. *admiración se atenúa* 'the admiration falls'). The second main class groups together collocations expressing five different PHASES of an emotion, such as PREPARATION (e.g. *buscar el afecto de alguien* 'aim to get sb's affection'), BEGINNING (e.g. *nacimiento de la amistad* 'the birth of a friendship'), DURATION (e.g. *cultivar la pasión* 'cultivate the passion'), REITERATION (e.g. *resucitar el rencor* lit. 'revive the grudge') and the END (e.g. *la alegría se desvanece* 'happiness vanishes'). The third and fourth semantic classes contain combinations whose meaning implies the MANIFESTATION (e.g. *mostrar amistad* 'show friendship') or the LACK OF MANIFESTATION (e.g. *tristeza reprimida* lit. 'repressed sadness') of an emotion. The fifth class named EXPERIENCER includes combinations whose meaning focuses on the first semantic actant of the emotion designated by the base of the collocation, i.e. the person experiencing it. These are typically support verb constructions such as *albergar esperanza* 'harbour hope', with the approximate meaning 'feel', while combinations such as e.g. *lleno de miedo* 'full of fear' also belong to this class. The sixth semantic class contains collocations expressing CAUSATION or making reference to the OBJECT of an emotion (e.g. *matar a alguien de aburrimiento* lit. 'kill sb of boredom', *gozar de la confianza de alguien* lit. 'enjoy sb's trust'). Finally, the seventh class includes combinations expressing QUALIFICATION (e.g. *simpatía mutua* 'mutual sympathy'), which can be further specified as being POSITIVE (e.g.

agradecimiento cálido lit. ‘warm gratitude’) or NEGATIVE nature (e.g. *atizar el ánimo* lit. ‘stir up the spirits’).

		Lexical Function	Semantic gloss	Collocation
Intensity	**High**	MagnOper$_1$, Magn+CausOper$_1$, Magn, Magn+Fact$_1$, etc.	feel an intense ~, make sb feel an intense ~, intense, ~ determines sb’s actions	*rebosar bienestar* ‘enjoy well-being’, *matar a alguien de aburrimiento* lit. ‘kill sb of boredom’, *loca alegría* lit. ‘crazy happiness’, *el deseo domina a alguien* lit. ‘desire dominates sb’
	Growing	CausPredPlus, IncepPredPlus, A$_1$IncepPredPlus, etc.	cause ~ to grow, ~ starts to grow, growing	*fortalecer una amistad* ‘strengthen a friendship’, *la amistad crece* ‘the friendship grows’, *agitación creciente* ‘growing anxiety’
	Low	AntiMagn, AntiMagn+Magn_temp, SAntiMagn, AntiMagn+A$_1$	low intensity, low intensity but continuous and lasting, having low ~	*escasa consideración* ‘little consideration’, *dolor sordo* ‘dull pain’, *bajo de ánimo* ‘have low spirits’
	Decreasing	CausPredMinus, IncepPredMinus	cause ~ to lessen, ~ becomes less intense	*apagar el deseo* ‘appease sb’s desire’, *la admiración se atenúa* ‘the admiration falls’
Phases	**Preparation**	PreparReal$_1$, Non-standardReal$_2$	propose a ~, aim to get ~	*ofrecer amistad a alguien* ‘offer one’s friendship to sb’, *buscar el afecto de alguien* ‘aim to get sb’s affection’
	Beginning	IncepFunc$_1$, SIncepFunc$_1$, Magn+IncepFunc$_1$	~ starts to be felt, beginning of ~, an intense ~ starts to exist in somebody	*le entran ganas a alguien* ‘start to feel like doing sg’, *nacimiento de la amistad* ‘the birth of a friendship’, *la tristeza embarga a alguien* ‘sadness seizes sb’
	Duration	ContOper$_1$, CausContFunc, Magn_temp	continue feeling ~, cause ~ to continue existing, lasting	*conservar la alegría* ‘retain happiness’, *cultivar la pasión* ‘cultivate the passion’, *antigua enemistad* ‘ancient enmity’
	Reiteration	Non-standardOper$_1$, Non-standardCausFunc	have ~ with sb again, make ~ exist again	*reanudar una amistad* ‘rekindle a freindship’, *resucitar el rencor* lit. ‘revive the grudge’
	End	FinOper$_1$, LiquOper$_1$, FinFunc$_1$	stop feeling ~, cause sb to stop having ~, ~ disappears	*perder la alegría* lit. ‘lose happiness’, *absolver alguien de la culpa* ‘absolve sb from the blame’, *la alegría se desvanece* ‘happiness vanishes’
Manifestation		Caus$_1$Manif, Adv$_1$Caus$_1$Manif, IncepManif	show ~, in order to show ~, ~ starts to show	*mostrar amistad* ‘show friendship’, *en señal de cariño* ‘as a sign of affection’, *la melancolía aflora* ‘melancholy erupts’

		Lexical Function	Semantic gloss	Collocation
Lack of manifestation		nonPerm$_1$Manif, A$_2$nonPerm$_1$Manif, SnonCaus$_1$Manif, etc.	make ~ not to show, ~ that does not show, absence of ~	*ocultar la alegría* 'hide happiness', *tristeza reprimida* lit. 'repressed sadness', *falta de interés* 'lack of interest'
Experiencer		Oper$_1$, nonOper$_1$, Func$_1$, Magn+A$_1$, etc.	feel ~, the ~ exists	*albergar esperanza* 'harbour hope', *la amistad une a alguien con alguien* 'friendship unites sb with sb', *lleno de miedo* 'full of fear'
Cause / Object		Caus$_2$Func$_1$, Magn+CuasOper1, Oper$_2$, etc.	cause sb's ~ to be directed to oneself, make sb feel an intense ~, be the object of ~	*ganar admiración* 'earn sb's admiration', *matar a alguien de aburrimiento* lit. 'kill sb of boredom', *gozar de la confianza de alguien* lit. 'enjoy sb's trust'
Qualification	Positive	Ver, Bon, MagnBon, CausPredBon	good, very good, make ~ better	*agradecimiento cálido* lit. 'warm gratitude', *gusto exquisito* 'exuisite taste', *levanter el ánimo* 'raise the spirits'
	Negative	AntiVer, AntiBon, CausPredAntiBon	unjustified, unfit, make ~ worse	*esperanza ilusa* 'unrealistic hope', *consuelo inútil* lit. 'useless consolation' *atizar el ánimo* lit. 'stir up the spirits'
	Other	Magn_quant$_1$, non-StandardA	shared by many, that is felt for one another	*aburrimiento generalizado* 'general boredom', *simpatía mutua* 'mutual sympathy'

Table 39 Most recurrent semantic categories identified in DiCE

In order to illustrate the alternative grouping of combinations made possible by the semantic classification presented in Table 39, we can compare the way collocations are currently presented in a given lexical entry of DiCE, and the way they can be reorganized using semantic classes. Table 40 shows the collocates expressing the approximate meanings 'intense' (underlined), 'beginning' (cursive) and 'end' (bold) as they are organized in DiCE, where combinations are primarily grouped according to their syntactic pattern in the entry of the base, with collocations corresponding to given LF forming subgroups. In comparison, Table 41 illustrates how the same collocations can be regrouped according to semantic classes. Note that in this latter case, a given collocation can belong to more than one class, such that, for instance, *el miedo invade* [*a alguien*] lit. 'fear invades [somebody]' and *el miedo asalta* [*a alguien*] lit. 'fear strikes [somebody]' appear both in the classes HIGH INTENSITY and BEGINNING.

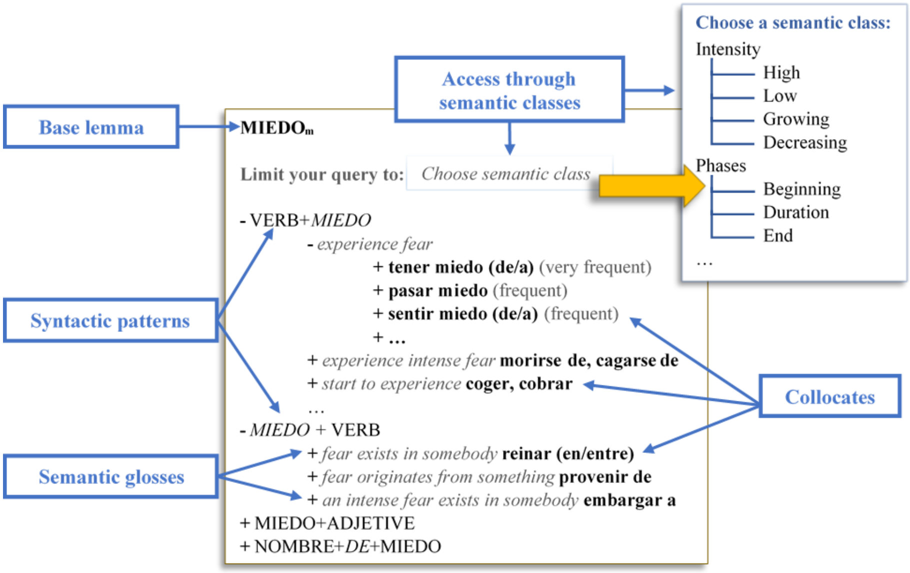
Participant attributes	‘who feels intense fear’ **muerto de miedo, lleno de miedo, etc.** …
Noun+adjective	‘intense fear’ **atroz, cerval, abrumador, etc.** …
Verb+noun	‘feel intense fear’ **cagarse de ~, morirse de ~, etc.** ‘start to feel fear’ ***coger ~ a, cobrar ~ a,* etc.** ‘stop feeling fear’ **perder el ~, etc.** ‘make somebody feel intense fear’ **llenar de ~** ‘make fear disappear’ **vencer el ~, superar el ~, etc.** …
Noun+verb	‘fear starts to be felt’ ***surgir, entrar a,* etc.** ‘intense fear starts to be felt’ ***invadir a, asaltar a,* etc.** ‘fear disappears’ **desvanecer, apagarse** ‘intense fear determines sb’s actions’ **dominar a** ‘intense fear is felt by many’ **cundir**

Table 40 Illustration of the current classification of collocations in DiCE

INTENSITY	HIGH	‘who feels intense fear’ **muerto de miedo, lleno de miedo, etc.** ‘intense fear’ **atroz, cerval, abrumador, etc.** ‘feel intense fear’ **cagarse de ~, morirse de ~, etc.** ‘make somebody feel intense fear’ **llenar de ~** ‘intense fear starts to be felt’ ***invadir a, asaltar a,* etc.** ‘intense fear determines sb’s actions’ **dominar a** ‘intense fear is felt by many’ **cundir**
	…	…
PHASE	BEGINNING	‘start to feel fear’ ***coger ~ a, cobrar ~ a,* etc.** ‘fear starts to be felt’ ***surgir, entrar a,* etc.** ‘intense fear starts to be felt’ ***invadir a, asaltar a,* etc.**
	END	‘stop feeling fear’ **perder el ~, etc.** ‘make fear disappear’ **vencer el ~, superar el ~, etc.** ‘fear disappears’ **desvanecer, apagarse**
	…	…

Table 41 Illustration of the novel grouping of collocations using semantic classes

It needs to be emphasized that the set of semantic classes described above should be understood as a mere starting point for a more comprehensive semantic taxonomy which aims to account for most combinations included in DiCE. In relation to this, the question arises as to whether it is possible to create a more comprehensive general taxonomy covering different semantic fields, which at the same time could be interpreted by, and hence would result useful for dictionary users. In this respect, recall that the analysis of user interactions with the DiCE interface, carried out as part of the usability test described in Chapter 5, revealed that participants had difficulty in interpreting semantic glosses aimed at expressing the meaning of collocations (see 5.2.5.4.A).



MIEDO$_m$

Limit your query to: *Choose semantic class*

- VERB+*MIEDO*
  - *experience fear*
    + **tener miedo (de/a)** (very frequent)
    + **pasar miedo** (frequent)
    + **sentir miedo (de/a)** (frequent)
    + **...**
  + *experience intense fear* **morirse de, cagarse de**
  + *start to experience* **coger, cobrar**
  …
- *MIEDO* + VERB
  + *fear exists in somebody* **reinar (en/entre)**
  + *fear originates from something* **provenir de**
  + *an intense fear exists in somebody* **embargar a**

+ MIEDO+ADJETIVE
+ NOMBRE+*DE*+MIEDO

Choose a semantic class:

Intensity
- High
- Low
- Growing
- Decreasing

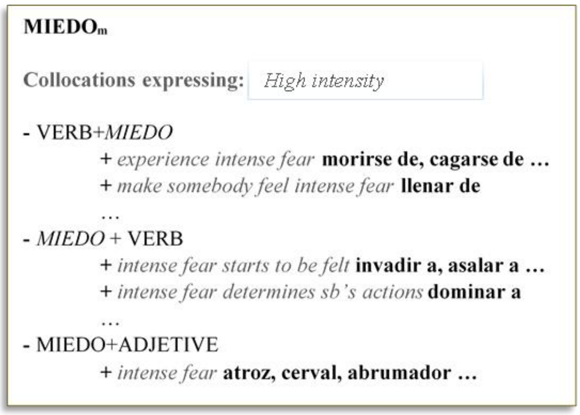
Phases
- Beginning
- Duration
- End

…

Figure 61 Sample collocation dictionary entry

Figure 61 shows a sample collocation dictionary entry which demonstrates how dictionary users can interact with semantic classes when searching the dictionary to browse collocations and potentially broaden their collocation repertoire. Note that in order to facilitate the interpretation of figures, all descriptive information, such as names of semantic classes, semantic glosses, etc. are shown in English throughout the present chapter, although in an actual tool targeted to SFL learners they would naturally figure in Spanish. Lexical information used to exemplify the dictionary module throughout this chapter comes from the DiCE database.

The dictionary entry shown here corresponds to the noun *miedo* ‘fear’, and would be displayed in the dictionary module when the user searches for this noun as a single lexical item (see Section 6.2.3 for a description of different possible search scenarios). The structure of the entry resembles that of a traditional collocation dictionary, with the exception that, at the top, the user is offered the option of filtering the combinations shown according to the semantic class they belong to. Since the user cannot be expected to be familiar with the labels applied for semantic classes, a pop-up window provides a list of the classes relevant in the given entry. Once the desired semantic class is chosen, the user is provided a list of corresponding combinations further classified according to their syntactic pattern. Figure 62 shows a sample dictionary entry where the semantic class filter was applied to limit the dictionary look-up to collocations belonging to the class HIGH INTENSITY.



MIEDO$_{\text{m}}$

Collocations expressing: *High intensity*

- VERB+*MIEDO*
    - + *experience intense fear* **morirse de, cagarse de …**
    - + *make somebody feel intense fear* **llenar de**
    - …
- *MIEDO* + VERB
    - + *intense fear starts to be felt* **invadir a, asalar a …**
    - + *intense fear determines sb's actions* **dominar a**
    - …
- MIEDO+ADJETIVE
    - + *intense fear* **atroz, cerval, abrumador …**

Figure 62 Collocations of the noun *miedo* expressing 'high intensity'

Some further features facilitating navigation in the lexical entry should also be mentioned in relation to Figure 61. Firstly, different types of information, in this case, collocation groups determined by syntactic pattern and semantic glosses can be contracted (-) or extended (+) if necessary, in order to adjust the amount of information displayed to the user's need. Secondly, when clicking on a specific collocation, users are given access to its corresponding lexical entry, providing more detailed information concerning the given combination. Criteria pertaining to the description of individual combinations, i.e. the content of collocation entries is discussed in the following section.

6.2.1.2 Describing collocations in the lexical entry

As we have seen in the case of the analysis of collocation errors in the learner corpus study presented in Chapter 4, erroneous lexical choice resulting in non-native-like combinations is not the only issue when it comes to language learners' collocation production. Grammatical errors constituting a deviation from the native-like collocation structure were also found to be common (see 4.5.3). What is more, it was also observed in the case of the error correction test using concordance data described in Chapter 5 that learners had more difficulty in correcting grammatical collocation errors than lexical errors when using authentic language data, suggesting that grammatical properties of combinations are less salient to learners (see 5.3.3.2). What follows from this is that, in order to constitute truly useful reference tools, collocation dictionaries cannot be confined to providing mere lists of co-occurring lexical items, but they need to provide a detailed description of the relevant characteristics of each combination. The same idea is expressed by Heid (2004), who, as it was already mentioned in 3.4.2.1.C, proposes what he refers to as a “maximalist” approach to the description of collocations, following which the lexical entry of a combination should provide 1) morphosyntactic, 2) syntactic, 3) semantic and 4) pragmatic information. In what follows, the presentation of each of these types of information in the collocation entry is discussed.

The prevalence of different types of grammatical collocation errors made by learners of Spanish identified in the learner corpus study described in 4.5.3, suggests that it is desirable to indicate in the lexical entry of a collocation morphosyntactic information including the gender of the noun(s) involved in the combination, as well as restrictions on the use of singular or plural forms, if they apply. Collocation errors observed in the learner corpus study affecting syntax involved the use of articles, governed prepositions, the government pattern of verbs, the use of pronominal verbs and, in the case of certain noun+adjective collocations, word order. Note that, part of this information, such as gender or government pattern can be found in e.g. monolingual learners’ dictionaries, however, its inclusion in a collocation dictionary, allows the learner to go without having to consult more than one reference tool in a given look-up situation.

Given the evidence concerning learner collocation errors as well as lack of salience in the input, I consider that the above mentioned morphosyntactic and syntactic features should be presented in the lexical entry of each collocation in an overt manner. In other words, they should be explicitly indicated either in the lemmatized form of the expression or explained in a usage note, instead of being implicitly present usage examples. Additionally, in order to avoid linguistic formalisms as well as the symbol “~”, often used to substitute the base when representing collocations, I suggest the use of “extended” lemmas, where all parts of the expression are spelt out similarly to the way shown in the example in (35). Since we are dealing with an electronic collocation dictionary, the inclusion of such description can be done without any concerns regarding space restrictions.

(35) **dejar** [a alguien] con las **ganas**$_{\mathbf{f}}$ de

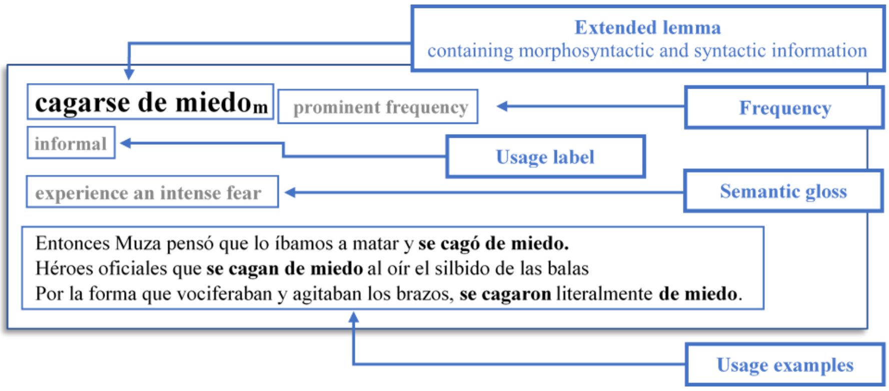
Regarding the description of the meaning of combinations in collocation dictionaries, 3.4.2.1.E mentioned a few usability studies (Komuro 2009; Lew and

Radłowska 2010), whose results highlighted the importance of the inclusion of explicit and easily interpretable semantic information. The previous section discussed the implementation of a comprehensive semantic typology used to classify collocations according to their meaning and independently of their syntactic pattern, with an aim of rendering dictionary access more flexible. However, it should be noted that semantic class labels associated with a collocation do not constitute a precise description of its meaning. As it was noted earlier, one collocation can belong to more than one semantic class, in addition, for the sake of generalizability, semantic class labels are rather abstract. In the corresponding lexical entry, the meaning of collocations can be indicated by a more specific semantic gloss, formulated in a way that it reflects the syntactic pattern of the combination, as it is done in the DiCE and some other combinatorial dictionaries (see 3.4.2.1.B and 3.4.2.1.C).

The last category Heid (2004) proposes to include in collocational lexical entries is pragmatic information, which refers to both usage labels and frequency information. As it was discussed in 4.3.2.1.B, the collocation error typology used for learner corpus annotation foresaw an error type concerning register errors, nevertheless, it was not productive in the specific dataset studied in Chapter 4, with only one instance of such error identified in the learner corpus. However, the inclusion of usage labels indicating pragmatic information is clearly useful for language learners, since it allows to identify combinations most suited to a given communicative purpose. As it was mentioned in 3.4.2.1.C, the potential of adding usage labels to the collocation entries was explored by Vázquez Veiga (2014) in the case of DiCE.

An indication of corpus frequency provides dictionary users with information regarding how commonly a given combination is used, and it can be especially useful when it comes to choosing a collocation from among a large number of near synonymous expressions. At the same time, frequency also constitutes useful information for language teachers and designers of teaching materials when selecting target combinations to be taught at different proficiency levels (see Martinez 2013; Nation 2001, 329). In order to present frequency information to the dictionary user in a more accessible way, instead of showing raw frequency values or scores corresponding to association measures, collocations can be assigned to frequency bands, such as e.g. *low, moderate, prominent, high* and *very high frequency*, which can be visually represented in the lexical entry through the use of colors or symbols. A methodology for calculating frequency scores to the collocations included in DiCE based on an annotated reference corpus and is described in more detail in Vincze and Alonso Ramos (2013).

By way of summary, Figure 63 shows a sample lexical entry for the collocation *cagarse de miedo* lit. ‘shit oneself of fear’ ‘be gripped by fear’ containing all of the above discussed information categories – morphosyntactic, syntactic, semantic and pragmatic information –, together with usage examples coming from corpus.



cagarse de miedo$_{m}$ prominent frequency

informal

experience an intense fear

Entonces Muza pensó que lo íbamos a matar y **se cagó de miedo.**
Héroes oficiales que **se cagan de miedo** al oír el silbido de las balas
Por la forma que vociferaban y agitaban los brazos, **se cagaron** literalmente **de miedo**.

Figure 63 Sample collocation entry provided in the dictionary module

6.2.2 Collocation corpus module

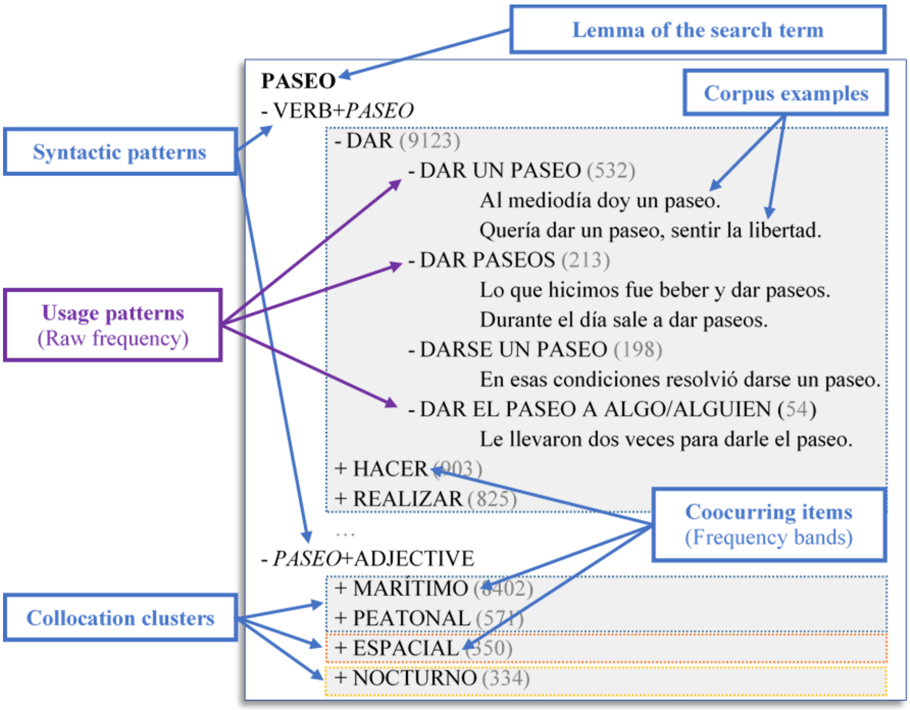
As it was proposed above, in a collocation learning tool integrating electronic dictionary and corpus searches, data coming from corpus can serve to complement information contained in the dictionary. Corpus data allows verifying whether a combination not included in the dictionary is attested and used frequently enough to be considered a native-like expression. A large corpus potentially provides access to a larger number of examples than a dictionary, and therefore can serve to explore usage patterns of a given combination or differences in the use of two similar combinations. While the previous subsection discussed how combinatorial information should be ideally organized and described in an electronic collocation dictionary, this subsection explores the types of information that can be obtained from language corpora, and the way corpus data can be organized and presented to best meet language learners’ needs.

A number of online collocation learning tools were reviewed in 3.4.2.3, among them *HARenEs*, a corpus tool in development designed specifically for learners of Spanish. We have seen that, depending on the query types available, these tools can provide lists of combinations containing a given lexical element (dictionary like tools) or a series of lexical elements (pattern search tools) as well as lists of combinations that have similar characteristics to the one the user introduced as a query term (collocation checkers and pattern search tools).

In order to automatically extract collocations from corpus, these are defined as strings of words corresponding to a series of given syntactic patterns or sequences of elements with specific parts of speech. Consequently, when displaying a list of combinations, they can be easily organized according to their syntactic pattern or the part of speech of composing elements. At the same time, similarly to the case of collocation dictionaries, it is also possible to group collocations according to their meaning. We have seen that some learning tools (e.g. *Collocation checker* and *Just the Word*) cluster collocations corresponding to the same syntactic pattern according to proximity of meaning. These clusters are usually created in an ad hoc manner and are unlabeled, however, some researchers have aimed at devising techniques that allow automatically sorting word combinations into pre-established classes corresponding to a semantic typology (for techniques concerning the automatic classification of collocations see e.g. Ferraro et al. 2011; Ferraro et al. 2014; Gelbukh and Kolesnikova 2013; Kolesnikova and Gelbukh 2010; Liu et al. 2009; Moreno et al. 2013; Nastase et al. 2006; J. C. Wu et al. 2010). While automatic semantic classification can be used to group the combinations containing a given lexical element in automatically generated collocation dictionary entries, it is also useful when it comes to collocation checking. Through identifying combinations with similar meanings, a collocation checker tool can suggest more appropriate alternative combinations to be used to substitute a non-native-like expression introduced by the learner.

In addition to syntactic pattern and semantic classification, a corpus-based tool can also provide information concerning the common usage patterns and frequency of collocations. One of the conclusions drawn from the study exploring language learners' ability in correcting collocation errors described in Chapter 5 was that grammatical features relevant in the use of collocations are not salient enough when presented implicitly in corpus examples (see 5.3.3.2). A way to overcome this problem is to display a list of frequent n-grams representing typical usage patterns under which corresponding corpus examples are grouped. This strategy, applied, as we have seen, in the case of the *Learning Collocations* module of *FLAX* (see 3.4.2.3.A), allows to sort concordance lines in a meaningful way, such that users can more easily infer the use of articles, prepositions, etc. as well as different uses or meanings of a given combination.

As in the case of the dictionary module, frequency information displayed together with combinations retrieved from the corpus provides an indication of how commonly a given expression is used. Frequency score should not only be shown in the case of combinations, but also individual usage patterns. Furthermore, combinations within collocation clusters, corresponding to groups of collocations with similar meanings, as well as n-grams, representing different usage patterns of a given combination, should be ordered from most frequent to least frequent to facilitate the interpretation of data. From among the tools more clearly oriented to language learners reviewed in 3.4.2.3, *FLAX*, *Just the Word*, *StringNet* and *Netspeak* provide information on raw frequency through displaying the number of occurrences found in the corpus along with combinations and/or usage patterns. Additionally, as we have seen, in the case of the search results offered by *Just the Word* a visual representation of the association strength characterizing the combination is also shown.



PASEO
- VERB+*PASEO*
  - DAR (9123)
    - DAR UN PASEO (532)
      Al mediodía doy un paseo.
      Quería dar un paseo, sentir la libertad.
    - DAR PASEOS (213)
      Lo que hicimos fue beber y dar paseos.
      Durante el día sale a dar paseos.
    - DARSE UN PASEO (198)
      En esas condiciones resolvió darse un paseo.
    - DAR EL PASEO A ALGO/ALGUIEN (54)
      Le llevaron dos veces para darle el paseo.
  + HACER (903)
  + REALIZAR (825)
  …
- *PASEO*+ADJECTIVE
  + MARÍTIMO (8402)
  + PEATONAL (571)
  + ESPACIAL (350)
  + NOCTURNO (334)

Figure 64 Sample lexical entry generated by the corpus module

Figure 64 shows a sample collocation dictionary entry generated from corpus data, which contains all the above described information. Such an entry would be displayed by the collocation corpus module when the search term introduced by the user consists of a single lexical item (different search scenarios are explained in detail below). Note that combinatorial information used to exemplify the collocation corpus module from here onwards represents hypothetical data, although it partially results from queries carried out using the *Sketch Engine* interface (Kilgarriff et al. 2004; Kilgarriff et al. 2014). In the specific case of this example, one particular drawback of corpus-generated combinatorial data can be examined, namely the lack of disambiguation of the different lexical units corresponding to the search term. This is apparent in the case of the adjectives *marítimo* 'maritime, coastal' and *peatonal* 'pedestrian', which are relevant for the meaning of the noun *paseo* 'path, public space destined for pedestrians' and not for the meaning 'walk'.

As it can be seen in Figure 64, similarly to the case of the dictionary module, in order to facilitate navigation on the screen, collocation groups determined by syntactic pattern, as well as information pertaining to a given combination or usage pattern can be contracted (-) or extended (+) if necessary. Since collocation clusters are not labeled, collocation groups belonging to a given cluster a visually marked in different colors.

6.2.3 Using the collocation learning tool: search scenarios

Following Verlinde et al. (2009, 8), access paths and feedback structures in a reference tool should be determined through reflecting on users' needs and defining possible tasks to be accomplished through dictionary searches. In the case of the tool proposed here, whose aim is to offer combinatorial information, the following five usage situations involving language production or reception can be anticipated.

Usage situations involving *production*:

a) The user wants to find out what other lexical items can a given lexical item be combined with to express a given meaning.
b) The user wants to know whether a given combination is native-like.
c) The user wants to find out about or verify the use of a given combination (use of the article, number, prepositions, etc.)
d) The user wants to verify whether the combinations in a text produced by them are native-like.

Usage situation involving *reception*:

e) The user wants to know or verify the meaning of a given combination.

With respect to the nature of feedback to be displayed, these five usage situations can be grouped under three main search scenarios: 1) In the case of a) the user introduces a single lexical item and expects to receive information on its combinatorial properties; 2) in the case of b), c) and e) the user requires information concerning a specific collocation; and 3) in the case of d), a user expects feedback on collocations found in a text, for which it is necessary to first identify collocations present in the text, and, second, retrieve information on the given combinations from the tool's database.

Although it is possible to conceive of implementing these search scenarios as different search options or search modules, results from the usability experiment conducted with the DiCE online interface suggest that this is not the most convenient solution. Recall that one of the observations made in this study was that dictionary users had difficulties using the highly modular search interface (see 5.2.5.4). Consequently, as it is further explained below, the collocation learning tool described here is proposed to contain a single search field where users can introduce single words, whole collocations or running text, while the feedback offered in each case is adapted to user needs corresponding to different usage situations.

The following section discusses a number of general features concerning the usability of the search interface. After this, I describe the search results or feedback provided to the user in the case of queries corresponding to each of the three search scenarios mentioned above: *single word search*, *collocation search* and *collocation verification in running text*.

6.2.3.1 Searching the interface

As mentioned above, results of the DiCE usability test showed that users frequently confused or misinterpreted the function of the individual query options (see 5.2.5.4 and 5.2.6). Consequently, it was suggested that the simplification of dictionary access would lead to more efficient dictionary consultation. The potential difficulty posed by modular searches was also noted by Lew (2012, 28), according to whom, users need a set of relevant skills to successfully use multiple search options available in an electronic dictionary. These include 1) recognizing the available access options, 2) selecting the one that best meets the user's information need, and 3) adapting the query in a way that it makes good sense at the given search option. In order to eliminate the need for such skills, the learning tool described here features a single multi-purpose search field, while search results or feedback are in each case adapted to users' perceived information need.

A similar solution is adopted in e.g. the *Interactive Language Toolbox*[41] (Serge Verlinde and Peeters 2012), whose users are presented with a single search box in which they can introduce single words, word combinations or a longer text. The authors of this

[41] https://ilt.kuleuven.be/inlato/

tool were inspired by de Schryver and Joffe's (2004) observation, according to which "users increasingly assume that electronic dictionaries behave like Web search engines such as *Google*, and type in concatenations of keywords, combinations and phrases surrounded by quotes, entire sentences, and even dump full paragraphs (lifted from other sources) into the search field." Similarly, the *Learning Collocations* module of the *FLAX* interface also offers a multi-purpose search field in which users can introduce either a single lexical item, which can constitute the base or the collocate in a collocation, or a whole combination (see Wu, Franken, and Witten 2010, 95).

While the introduction of an all-purpose search box releases users from having to pick the appropriate search option that suits their information need, the implementation of strategies such as *fuzzy-spelling search*, *incremental search* and, *inflected form search* aid them in formulating queries. All three of these techniques are aimed at helping dictionary users overcome the difficulty of introducing the search term correctly. The first one, fuzzy-spelling search (Lew 2013, 26; see also Lew and Mitton 2011; Lew and Mitton 2013), also known as 'did-you-mean' function, provides suggestions for potentially misspelled items introduced in the search field of the dictionary. In other words, it serves to compensate for users' spelling and/or typing errors.

A second technique that can contribute to resolving users' uncertainty regarding the spelling of words is incremental search. This feature involves the automated completion of the search term while it is being typed, through displaying matching terms from an index in a drop-down menu. The feature is bound to be familiar to the users of a learning tool, as it is offered by popular user interfaces, such as e.g. the *Wikipedia* or web search engines such as *Google* (see Lew 2013, 24; Lew 2012, 151–152). In the case of the collocation learning tool, incremental search has to allow access not only to single lexical items, but also to multiword combinations. This means that, while typing in the search field, the user is provided a list of matching single word items and collocations found in the database of the learning tool.

Finally, the implementation of inflected form search (Lew 2013, 26), which allows to obtain search results not only when introducing the conventional lemma form of a lexical item, but also when entering an inflected form, is necessary not only in order to overcome cases when the user is not familiar with a given citation form, but also to facilitate full collocation searches. The correct or the natural formulation of a collocation often includes lexical items that do not occur in the citation form. For instance, when searching for a noun+adjective combination containing a feminine gender noun (e.g. *sospecha ligera* ‘slight suspicion), it is unnatural/ungrammatical to introduce the collocate adjective in its singular masculine form. Similarly, in the case of noun+verb combinations (e.g. *el viento sopla* ‘the wind blows’), it may result unnatural to use the infinitive form of the verb in the query (*viento soplar* ‘wind blow’). In addition, the correct formulation of many collocations involves the use of a preposition or an article between the collocate and the base, or the use of the plural form of a noun. In order to carry out efficient full collocation searches, the user should be able to retrieve, for instance, the lexical entry corresponding to the combination with the base *gana*(*s*) and the collocate *morir* through introducing either of the following strings: *morir de ganas*, *muere de ganas morir de las ganas de *morir de la gana, *morir las ganas*.

6.2.3.2 Displaying search results

In order to better integrate the information offered by the dictionary and corpus modules presented above, search results coming from both sources can be displayed simultaneously on a shared screen. Such an arrangement allows the user to access and compare different types of information, similarly to the case of the online dictionary site *Wordnik*[42], which displays lexical data retrieved from different sources including dictionaries and web corpora[43].

As it was mentioned above, combinatory information coming from dictionary and corpus can be seen as complementing each other. Firstly, while the content of the dictionary module can be regarded as more reliable, existing lexicographical description of combinatorial information is complemented by the additional examples, usage patterns and combinations potentially offered by the corpus module. Furthermore, in the case of queries where the dictionary module does not contain relevant information as yet, the corpus module has the potential of supplying the combinations required by the user. Secondly, the description of combinatory information in the dictionary module compiled by lexicographers can aid the user in interpreting and filtering less structured and less reliable corpus data.

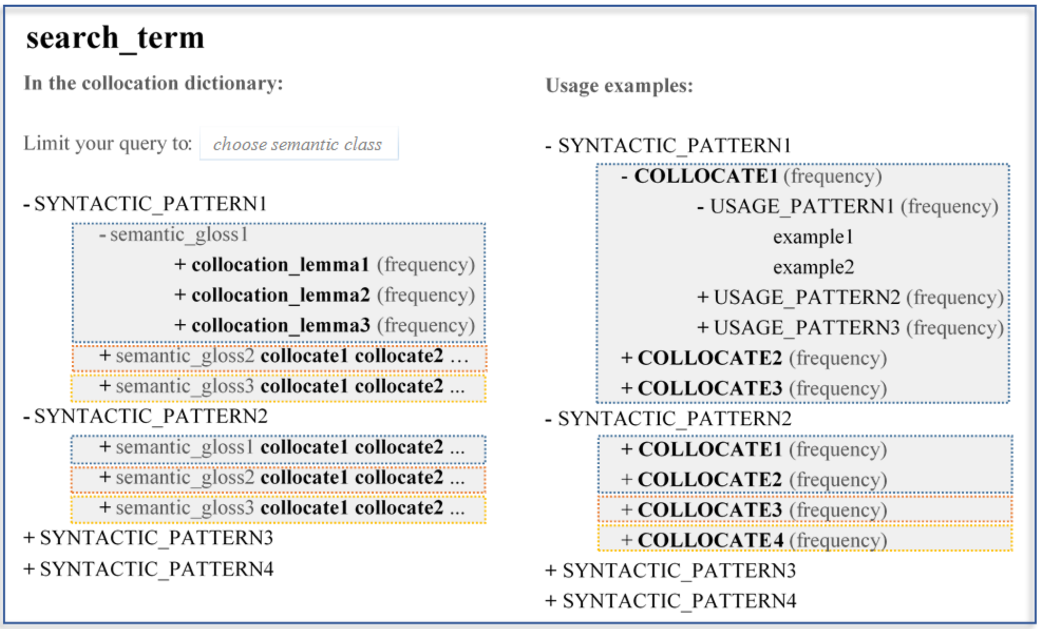
[42] www.wordnik.com

[43] Different types of information on words provided by *Wordnik* include definitions coming from dictionaries and thesauri, examples retrieved from different sources such as web corpora and the *Twitter* API, sound files, and images from *Fickr*.

In what follows, the search results or feedback displayed in the case of each of the three search scenarios foreseen are considered one by one. As explained above, these include searching for 1) the combinatorial properties of a lexical item, or 3) a specific collocation and 3) the verification of collocations in running text.

A. *Single word search*

When users introduce a single lexical item in the search field, it is assumed that they are aiming to find information on the combinatory characteristics of the given word, i.e. they expect to be provided a list of lexical items the search term can be combined with. This search scenario is analogous to look-up situations in the case of a collocation dictionary, which means that the search results provided should be similar to the lexical entries found in these. Criteria for organizing combinatorial information as well as for the description of collocations in both the dictionary and the corpus module were discussed above in detail, consequently here I merely illustrate the way information coming from both the dictionary and the corpus module are displayed simultaneously.



search_term

In the collocation dictionary:

Limit your query to: *choose semantic class*

- SYNTACTIC_PATTERN1

- semantic_gloss1

+ **collocation_lemma1** (frequency)

+ **collocation_lemma2** (frequency)

+ **collocation_lemma3** (frequency)

+ semantic_gloss2 **collocate1 collocate2** ...

+ semantic_gloss3 **collocate1 collocate2** ...

- SYNTACTIC_PATTERN2

+ semantic_gloss1 **collocate1 collocate2** ...

+ semantic_gloss2 **collocate1 collocate2** ...

+ semantic_gloss3 **collocate1 collocate2** ...

+ SYNTACTIC_PATTERN3

+ SYNTACTIC_PATTERN4

Usage examples:

- SYNTACTIC_PATTERN1

- **COLLOCATE1** (frequency)

- USAGE_PATTERN1 (frequency)

example1

example2

+ USAGE_PATTERN2 (frequency)

+ USAGE_PATTERN3 (frequency)

+ **COLLOCATE2** (frequency)

+ **COLLOCATE3** (frequency)

- SYNTACTIC_PATTERN2

+ **COLLOCATE1** (frequency)

+ **COLLOCATE2** (frequency)

+ **COLLOCATE3** (frequency)

+ **COLLOCATE4** (frequency)

+ SYNTACTIC_PATTERN3

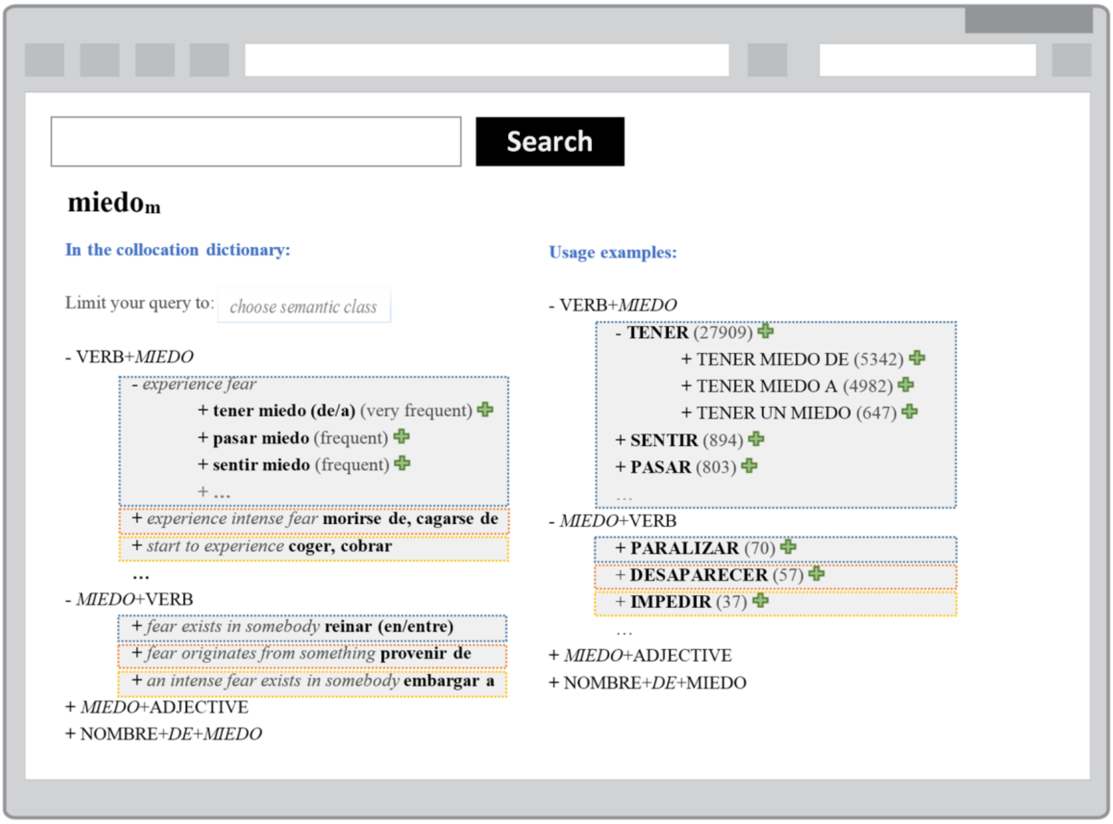
+ SYNTACTIC_PATTERN4

Figure 65 Presentation of search results in the “single word search” scenario

Figure 65 models the feedback provided to the user when searching for a single word. Information coming from the dictionary module is displayed on the left hand side (*collocation dictionary*), while data retrieved from the corpus module is shown on the right hand side (*usage examples*). As it was explained in 6.2.1, in the case of the dictionary module, the combinations displayed are grouped according to their syntactic pattern, and listed within each group together with a semantic label representing their meaning. At the same time, the user has the option of filtering the results according to semantic classes (see Figure 62). In the corpus module, as described in 6.2.2, combinations are grouped according to their syntactic pattern and approximate meaning. In the case of each particular collocation, examples are further grouped according to the usage pattern they represent with the aim of highlighting features such as government pattern, prepositions, articles, etc. As mentioned above, displaying combinatory information retrieved from the dictionary and corpus modules simultaneously allows the user to have access to two types of information, potentially complementing each other, at one stroke: more accurate and reliable data coming from a collocation dictionary, and corpus data that can potentially provide further examples as well as illustrate combinatorial phenomena not as of yet covered by the dictionary.

Note that, in Figure 65 co-occurring lexical items listed in the dictionary and the corpus modules are labeled as *collocate*, i.e. the search term is assumed to constitute the base of the collocation. This annotation, however, is used merely for the sake of simplifying the representation. In the case of single word searches, the results provided by the collocation learning tool list combinations including the search term both as base and collocate in equal detail. Recall that, as explained in 3.4.2.1.B, collocation dictionaries tend to deal differently with these two cases, and that it is assumed that listing the collocates in the lexical entry of the base suits language production better, while listing bases commonly occurring with a given collocate is assumed to benefit decoding.

Figure 66 models the search results displayed in the case of a single word query involving the noun *miedo* ‘fear’. Since nouns typically constitute the base in collocations, the information provided includes a list of possible collocates. In contrast, Figure 67 shows the combinatory information displayed as a result of a search for the lexical item *vencer* ‘defeat’, which, being a verb, more frequently constitutes the collocate and not the base in a collocation. In fact, in the case of the dictionary module, only collocations having *vencer* as a collocate are shown; this is so since the data set used here to exemplify the dictionary module comes from DiCE, only containing combinations with noun bases. On the contrary, in the corpus module, a group of combinations is shown where the verb constitutes the base, and combines with adverbial collocates. This serves as an example of how corpus data can complement the dictionary. Notice also that, while, in the case of the dictionary module, all collocations are grouped according to the meaning of the combination, in the case of the corpus module, it is possible to cluster combinations according to the meanings of possible bases. This way of presenting combinatorial information is analogous to what we have seen in the case of *Redes* (Bosque 2004b), see 3.4.2.1.B.



miedo$_{\mathbf{m}}$

In the collocation dictionary:

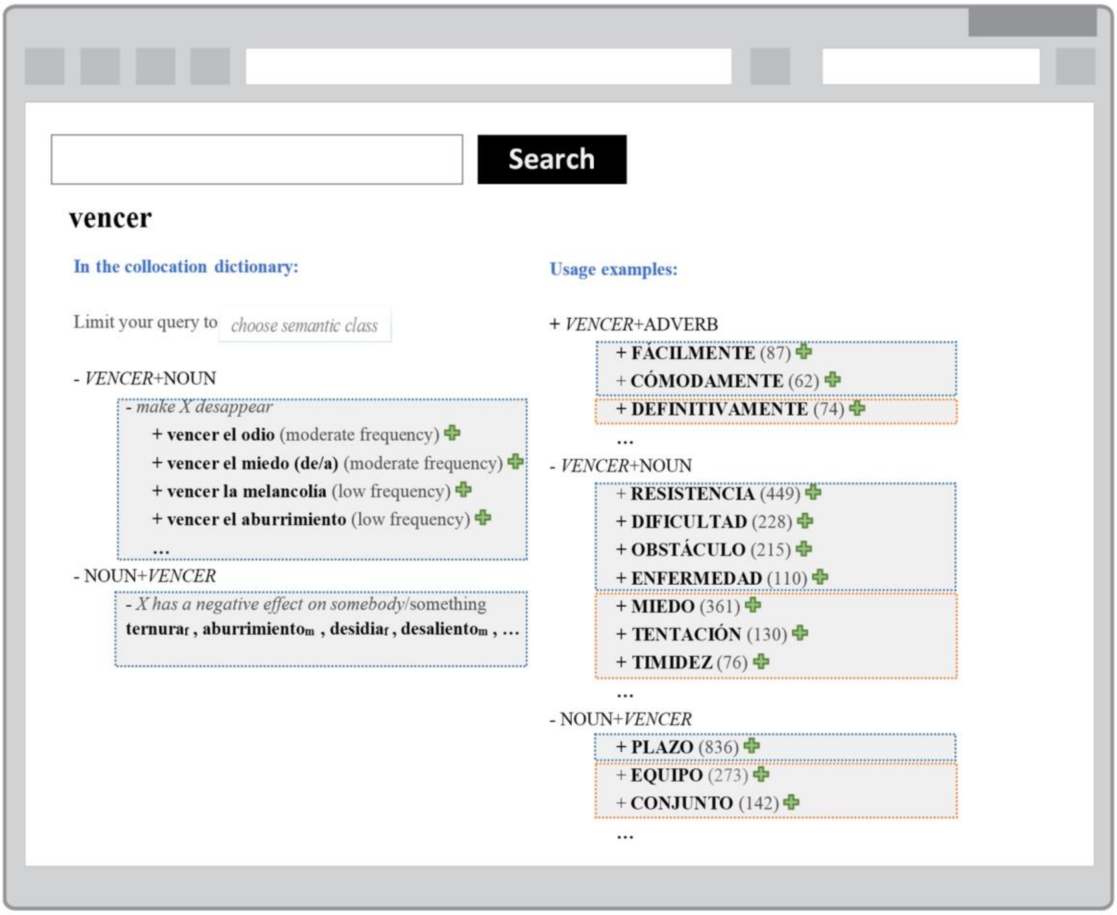
Limit your query to: *choose semantic class*

- VERB+*MIEDO*
  - *experience fear*
    + **tener miedo (de/a)** (very frequent) ✚
    + **pasar miedo** (frequent) ✚
    + **sentir miedo** (frequent) ✚
    + …
  + *experience intense fear* **morirse de, cagarse de**
  + *start to experience* **coger, cobrar**
  …
- *MIEDO*+VERB
  + *fear exists in somebody* **reinar (en/entre)**
  + *fear originates from something* **provenir de**
  + *an intense fear exists in somebody* **embargar a**
+ *MIEDO*+ADJECTIVE
+ NOMBRE+*DE*+*MIEDO*

Usage examples:

- VERB+*MIEDO*
  - **TENER** (27909) ✚
    + TENER MIEDO DE (5342) ✚
    + TENER MIEDO A (4982) ✚
    + TENER UN MIEDO (647) ✚
  + **SENTIR** (894) ✚
  + **PASAR** (803) ✚
  …
- *MIEDO*+VERB
  + **PARALIZAR** (70) ✚
  + **DESAPARECER** (57) ✚
  + **IMPEDIR** (37) ✚
  …
+ *MIEDO*+ADJECTIVE
+ NOMBRE+*DE*+MIEDO

Figure 66 Search results for the single word query involving the noun *miedo* ‘fear’



vencer

In the collocation dictionary:

Limit your query to *choose semantic class*

- *VENCER*+NOUN

  - *make X desappear*
    - + **vencer el odio** (moderate frequency) ✚
    - + **vencer el miedo (de/a)** (moderate frequency) ✚
    - + **vencer la melancolía** (low frequency) ✚
    - + **vencer el aburrimiento** (low frequency) ✚
    - …

- NOUN+*VENCER*

  - *X has a negative effect on somebody*/something

    ternura$_f$, aburrimiento$_m$, desidia$_f$, desaliento$_m$, …

Usage examples:

+ *VENCER*+ADVERB

  + **FÁCILMENTE** (87) ✚
  + **CÓMODAMENTE** (62) ✚
  + **DEFINITIVAMENTE** (74) ✚
  …

- *VENCER*+NOUN

  + **RESISTENCIA** (449) ✚
  + **DIFICULTAD** (228) ✚
  + **OBSTÁCULO** (215) ✚
  + **ENFERMEDAD** (110) ✚
  + **MIEDO** (361) ✚
  + **TENTACIÓN** (130) ✚
  + **TIMIDEZ** (76) ✚
  …

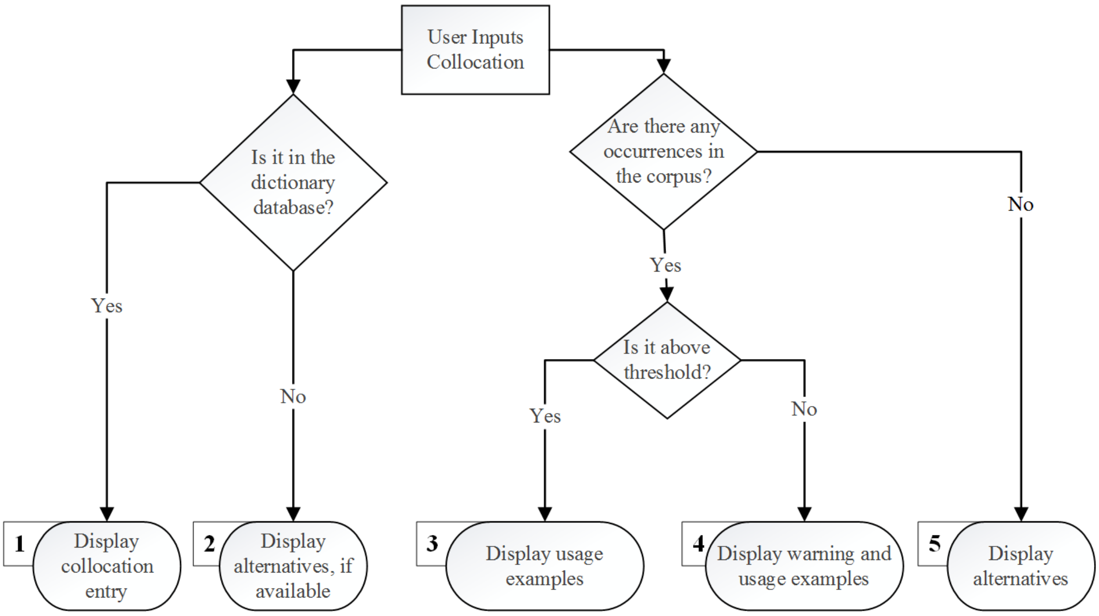
- NOUN+*VENCER*

  + **PLAZO** (836) ✚
  + **EQUIPO** (273) ✚
  + **CONJUNTO** (142) ✚
  …

Figure 67 Search results for the single word query involving the verb *vencer* ‘defeat’

B. Collocation search

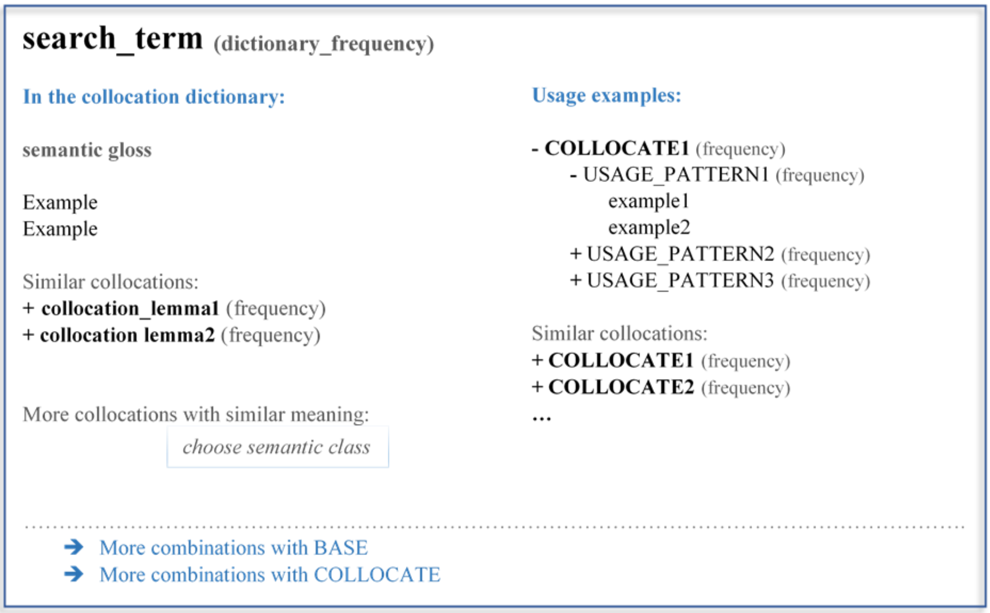
It can be assumed that users introducing a whole collocation in the search field either want to learn about the meaning or the use of a given combination, or they aim to verify whether the given combination is native-like. While in the case of the single word search scenario, discussed above, the search results offered by the learning tool are largely analogous to lexical entries found in collocation dictionaries, in the case of full collocation search, as shown in Figure 68, the nature of the information displayed by the tool as a result of a query varies depending on 1) whether the collocation introduced by the learner has a corresponding entry in the dictionary database, 2) whether occurrences of it can be found in the corpus, and 3) whether it can be considered as a typical or native-like combination based on given criteria. In what follows the information displayed in each of these search situations is exemplified in more detail.


User Inputs Collocation
Is it in the dictionary database?
Yes
No
Are there any occurrences in the corpus?
Yes
No
Is it above threshold?
Yes
No
1 Display collocation entry
2 Display alternatives, if available
3 Display usage examples
4 Display warning and usage examples
5 Display alternatives

1	2	3	4	5
See Figure 70	See Figure 71 and Figure 72	See Figure 70	See Figure 71	See Figure 72

Figure 68 Summary of different types of feedback provided in the "search collocations" scenario

Following Figure 68, when the combination introduced by the user can be found in the dictionary database, its corresponding collocation entry is displayed as a result of the search, while usage patterns and examples retrieved from the corpus are also shown when available (see Figure 69).



search_term (dictionary_frequency)

In the collocation dictionary:

semantic gloss

Example
Example

Similar collocations:
+ **collocation_lemma1** (frequency)
+ **collocation lemma2** (frequency)

More collocations with similar meaning:
choose semantic class

➔ More combinations with BASE
➔ More combinations with COLLOCATE

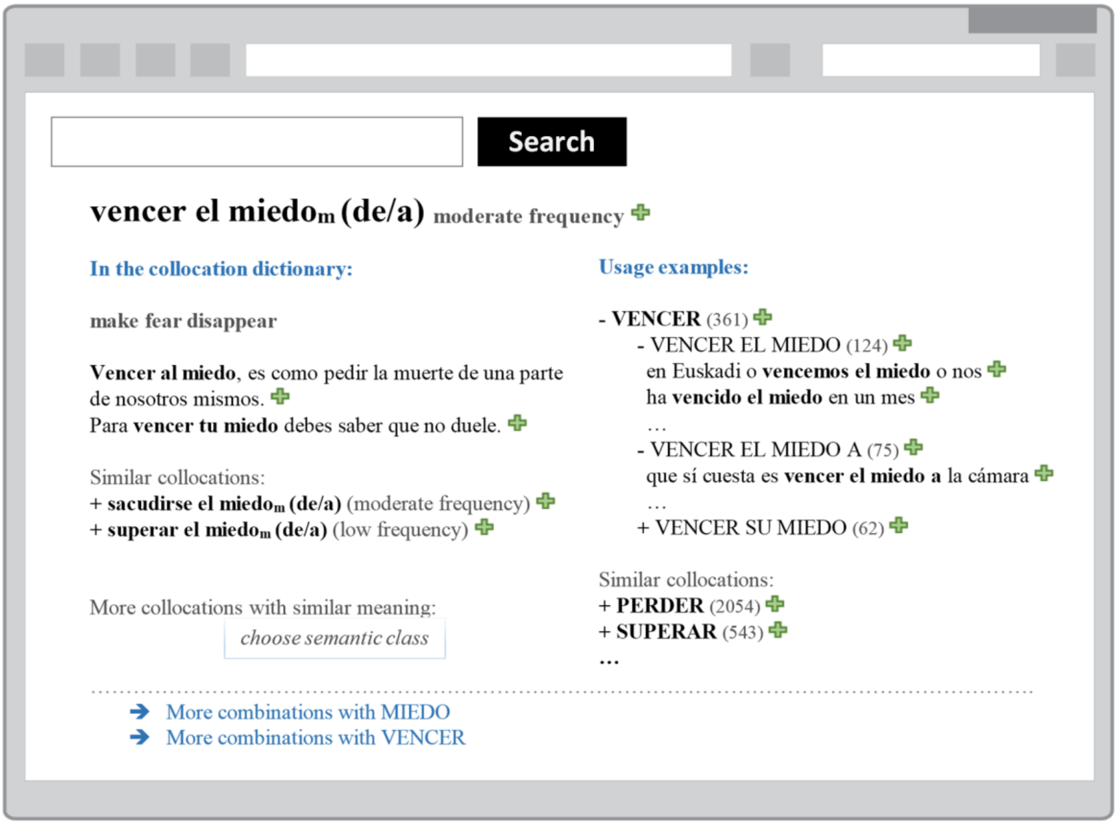
Usage examples:

- **COLLOCATE1** (frequency)
    - USAGE_PATTERN1 (frequency)
        example1
        example2
    + USAGE_PATTERN2 (frequency)
    + USAGE_PATTERN3 (frequency)

Similar collocations:
+ **COLLOCATE1** (frequency)
+ **COLLOCATE2** (frequency)
…

Figure 69 Presentation of search results in the "collocation search" scenario

As it can be seen in Figure 69, the information provided on specific combinations aims to satisfy user needs concerning both the meaning and the use of the collocation. In addition, the dictionary module also provides direct access to combinations related to the search term, such as (near) synonymous combinations having the same syntactic pattern and semantic gloss (*Similar collocations*) and collocations belonging to the same semantic classes (*More collocations with similar meaning*). In the case of the corpus module, besides the usage patterns corresponding to the collocation introduced in the query, *similar collocations*, i.e. combinations having the same base and belonging to the same collocation cluster (see below) as the search term are also displayed (*Similar collocations*).



vencer el miedo$_m$ (de/a) moderate frequency ✚

In the collocation dictionary:

make fear disappear

Vencer al miedo, es como pedir la muerte de una parte de nosotros mismos. ✚
Para **vencer tu miedo** debes saber que no duele. ✚

Similar collocations:
+ **sacudirse el miedo$_m$ (de/a)** (moderate frequency) ✚
+ **superar el miedo$_m$ (de/a)** (low frequency) ✚

More collocations with similar meaning:
choose semantic class

Usage examples:

- **VENCER** (361) ✚
  - VENCER EL MIEDO (124) ✚
    en Euskadi o **vencemos el miedo** o nos ✚
    ha **vencido el miedo** en un mes ✚
    …
  - VENCER EL MIEDO A (75) ✚
    que sí cuesta es **vencer el miedo a** la cámara ✚
    …
  + VENCER SU MIEDO (62) ✚

Similar collocations:
+ **PERDER** (2054) ✚
+ **SUPERAR** (543) ✚
…

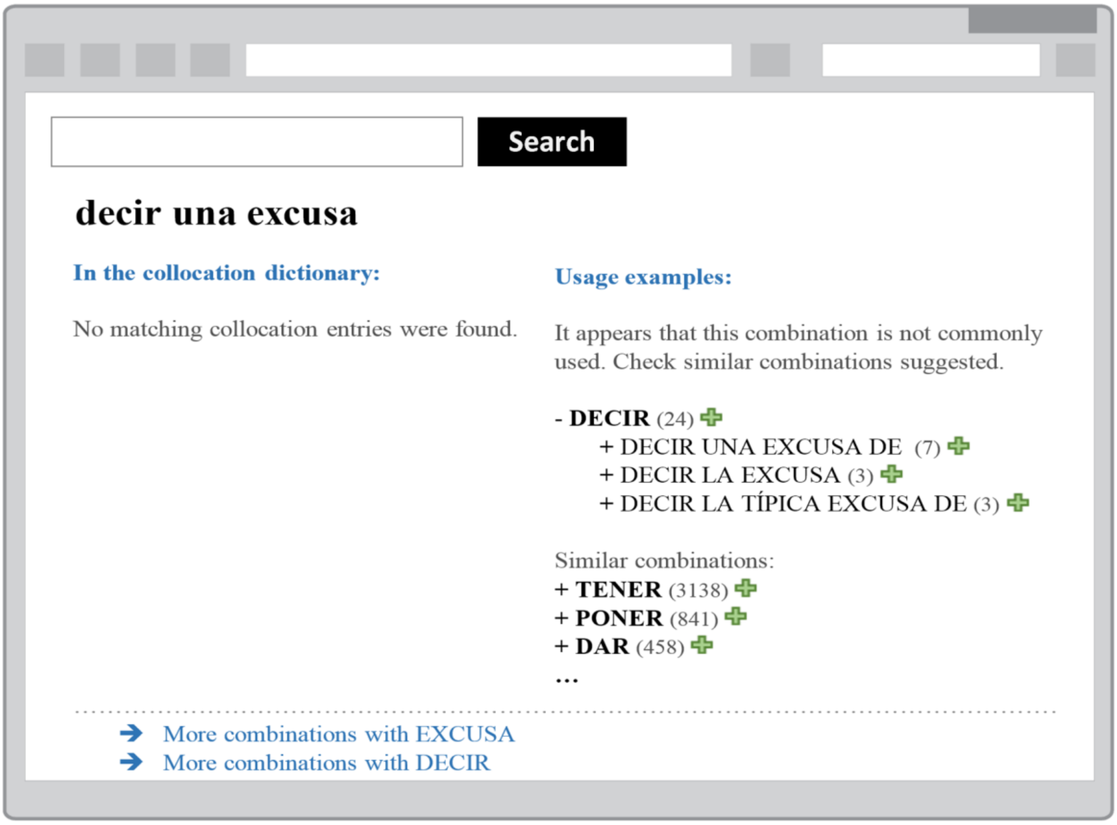
➔ More combinations with MIEDO
➔ More combinations with VENCER

Figure 70 Search results for the query involving the collocation *vencer el miedo* ‘overcome fear’

When the aim of a query is to verify whether a given combination is used in Spanish, and the search results show that there is a matching collocation entry in the dictionary database, the user can conclude that the search term introduced constitutes a native-like combination, as is the case of the collocation *vencer el miedo* ‘overcome one’s fear’ (see Figure 70). However, the fact that a combination introduced by the learner does not have a lexical entry in the collocation dictionary naturally does not necessarily mean that it is not native-like; it can be simply missing from the dictionary nomenclature. This situation should be familiar to learners, given it is not uncommon in the case of general monolingual or bilingual learner dictionaries. Importantly, in the case of the collocation learning tool described here, when a given collocation entry is lacking in the dictionary, through looking at the information offered by the corpus module, users have the option to decide whether they will use the combination or not.

While users of the collocation learning tool can be expected to make judgments as to the appropriateness of a given combination through relying on the information provided by the corpus module, giving automatic explicit feedback regarding the correctness of collocations constitutes an appealing solution in helping learners decide on which combinations to use. However, as it was discussed in 3.4.2.3.C, feedback consisting of straightforwardly stating whether a combination is correct or not may result misleading to the language learner unaware of the limitations of state of the art technology. This is so since, while the output of the learning tool may be prone to error, the fact that there is straightforward feedback may give the user the false impression of interacting with an online tutor (see Milton and Cheng 2010). This is the reason why the learning tool described here offers a more cautious formulation of feedback, combined with information concerning the corpus frequency of the query term and suggested alternative combinations. Consequently, when a combination queried by the user scores below a given threshold value established on the basis of corpus data, instead of stating that the search term is incorrect, the tool provides a warning and encourages the learner to look at a list of similar but more typical combinations, see Figure 71.

As it was suggested in 4.5.3, since lexical errors can affect both the base and the collocate, a collocation checker tool should provide alternative combinations both through substituting the collocate and the base in the expression constituting the search term. For instance, in the case of the combination ?*decir una excusa* lit. ‘say an excuse’, which, as shown in Figure 71, does not reach the established threshold in the corpus, the tool should provide a list of combinations in which the noun *excusa* co-occurs with other possible verbal collocates, and combinations of verb *decir* with other nouns whose meaning is potentially similar to that of *excusa*. As it was explained in 3.4.2.3.C, this strategy is applied in the *Alternatives* module of *Just the Word*. Nevertheless, since the study of learner collocation errors, described in Chapter 4, showed that the majority of lexical errors concern the collocate (see 4.4 and 4.5.3), in the case of the learning tool described here, I opted for displaying only a list of alternative combinations containing the same base in the main area dedicated to search results. The user can access further alternatives through the links provided underneath the search results which retrieve information corresponding to a single word search carried out for the base or the collocate respectively. For instance, in the case of the whole collocation search involving the combination ?*decir una excusa*, shown in Figure 71, the user can click on *More combinations with EXCUSA* and *More combinations with DECIR*, in order to launch a single word query for the base or the collocate respectively.



decir una excusa

In the collocation dictionary:

No matching collocation entries were found.

Usage examples:

It appears that this combination is not commonly used. Check similar combinations suggested.

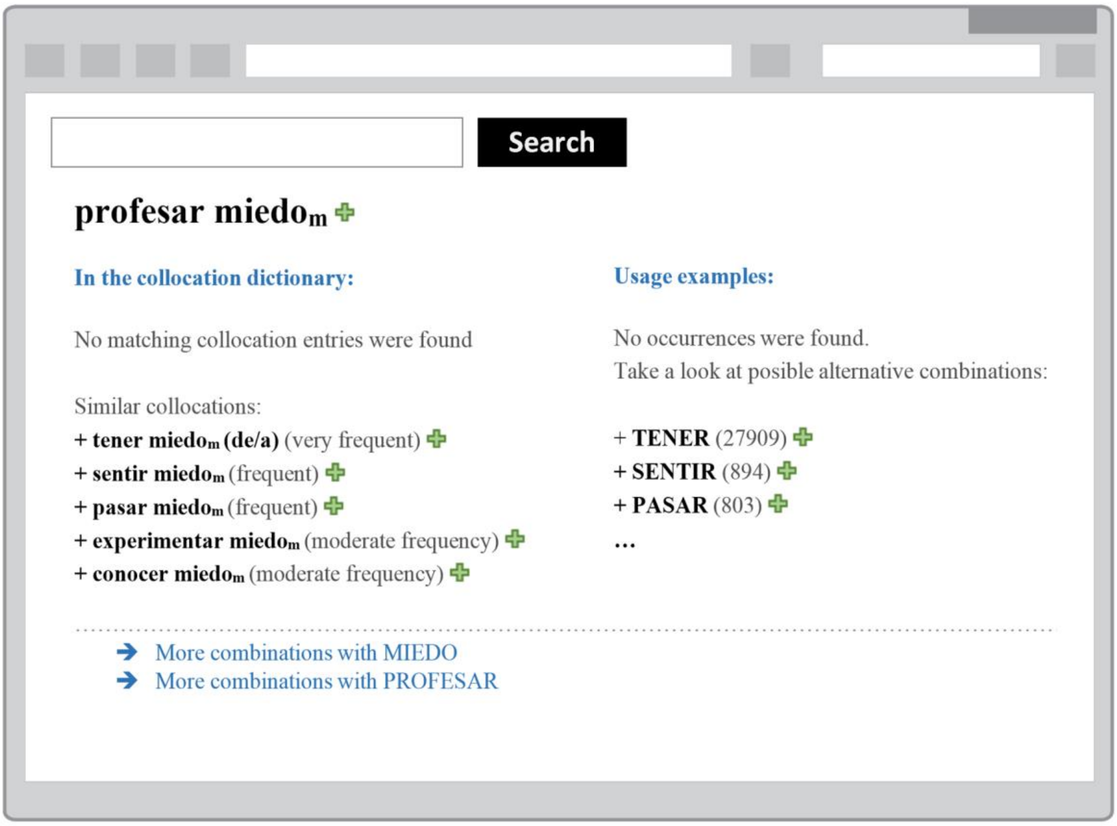
- **DECIR** (24) ✚
  + DECIR UNA EXCUSA DE (7) ✚
  + DECIR LA EXCUSA (3) ✚
  + DECIR LA TÍPICA EXCUSA DE (3) ✚

Similar combinations:
+ **TENER** (3138) ✚
+ **PONER** (841) ✚
+ **DAR** (458) ✚
...

➔ More combinations with EXCUSA
➔ More combinations with DECIR

Figure 71 Suggestion of alternative collocations to be used instead of an infrequent combination in the corpus module, exemplified in the case of the combination *decir una excusa* 'say an excuse'

As it was mentioned above, in the case of the corpus module, collocations with an assumedly similar meaning to that of the search term can be offered through displaying combinations belonging to the same collocation cluster. This essentially means displaying a set of potentially synonymous collocates that can co-occur with the base to form a collocation having the same syntactic pattern. The same strategy can be also applied when no occurrences of the combination introduced by a user can be found in the database corresponding to the corpus module, as in the case of the combination **profesar miedo* shown in Figure 72. Techniques that allow to retrieve semantically similar candidate collocates and, consequently, to offer correction suggestions in the case of inappropriate combinations are described in e.g. Chang et al. (2008), Dahlmeier and Ng (2011), Ferraro et al. (2014), Kochmar and Briscoe (2015) and Liu et al. (2009).



profesar miedo$_{m}$ ✚

In the collocation dictionary:

No matching collocation entries were found

Similar collocations:

- + **tener miedo$_{m}$ (de/a)** (very frequent) ✚
- + **sentir miedo$_{m}$** (frequent) ✚
- + **pasar miedo$_{m}$** (frequent) ✚
- + **experimentar miedo$_{m}$** (moderate frequency) ✚
- + **conocer miedo$_{m}$** (moderate frequency) ✚

Usage examples:

No occurrences were found.
Take a look at posible alternative combinations:

- + **TENER** (27909) ✚
- + **SENTIR** (894) ✚
- + **PASAR** (803) ✚

…

➔ More combinations with MIEDO
➔ More combinations with PROFESAR

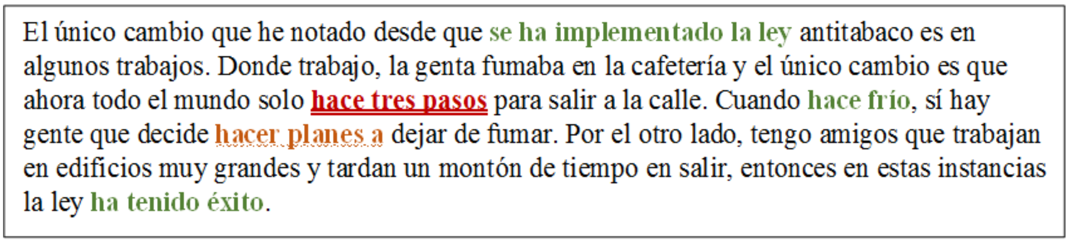
Figure 72 Presentation of search results in the "collocation search" scenario with no matching dictionary entries and no occurrences in the corpus

As for the dictionary module, note that in the case of Figure 71 showing hypothetical data representing feedback for the combination ?*decir una excusa* 'say an excuse' input by the user, no data whatsoever is offered. This is so because there is no lexical entry for the given combination, and the dictionary database does not contain any collocations with the base *excusa*. In other cases, when the base of the combination queried by the user is included in the dictionary nomenclature, it is possible to offer collocation alternatives through making use of LFs. Recall that LFs encode syntactic and semantic information simultaneously, so that two collocations represented by the same LF will have the same syntactic configuration and the same generic meaning. Consequently, when the learner searches for a collocation not included in the dictionary, such as e.g. the erroneous combination **profesar miedo* 'to feel fear', it is possible to obtain a list of alternative collocations through looking up what LF is associated with collocations containing the collocate verb *profesar* in the dictionary database, in this case Oper$_1$, and then listing all collocations of the base *miedo* described by the same LF, such as *tener miedo, sentir miedo, pasar miedo*, etc. (see Figure 72).

C. Collocation verification in running text

As we have seen in the case of *Collocation Inspector* (Chang et al. 2008), a tool described in 3.4.2.3.C, collocation checking or collocation verification can be implemented as a writing aid tool, allowing the quick identification of potentially problematic combinations in running text. According to Verlinde et al. (2009, 12), writing assistants are especially useful when they allow users to check their texts on features not covered by regular spelling and grammar checkers included in word processors, which is precisely the case of a tool capable of identifying collocation errors.

Collocations identified in learner texts can be verified through contrasting them with combinations in the dictionary database and those occurring in the corpus. The aim of the tool is not only to detect potential errors, but also to help the language learner to explore new collocations and use the most appropriate combination in a given context. Therefore, in addition to offering feedback regarding the correction of combinations, the tool also provides direct access to information concerning each of the collocations found in the learner text, regardless of whether they are judged to be native-like or not. This means that, as a result of the verification process, all collocations in the users' text are highlighted, in a way that combinations attested in the dictionary and/or the corpus can be clearly differentiated from unattested combinations and combinations below threshold (see Figure 73). In order to proceed to the revision of their text, users can click on highlighted collocations to receive more information.



El único cambio que he notado desde que se ha implementado la ley antitabaco es en algunos trabajos. Donde trabajo, la genta fumaba en la cafetería y el único cambio es que ahora todo el mundo solo hace tres pasos para salir a la calle. Cuando hace frío, sí hay gente que decide hacer planes a dejar de fumar. Por el otro lado, tengo amigos que trabajan en edificios muy grandes y tardan un montón de tiempo en salir, entonces en estas instancias la ley ha tenido éxito.

Figure 73 Sample output of "collocation verification" with correct and incorrect collocations highlighted in the learner text

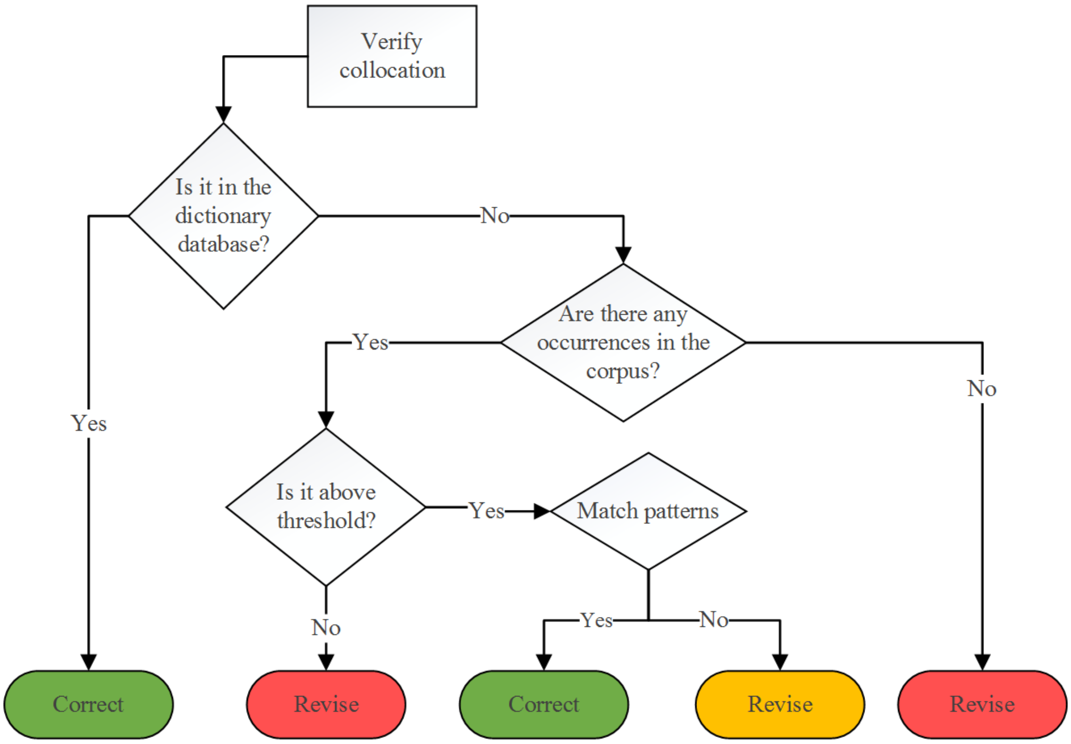
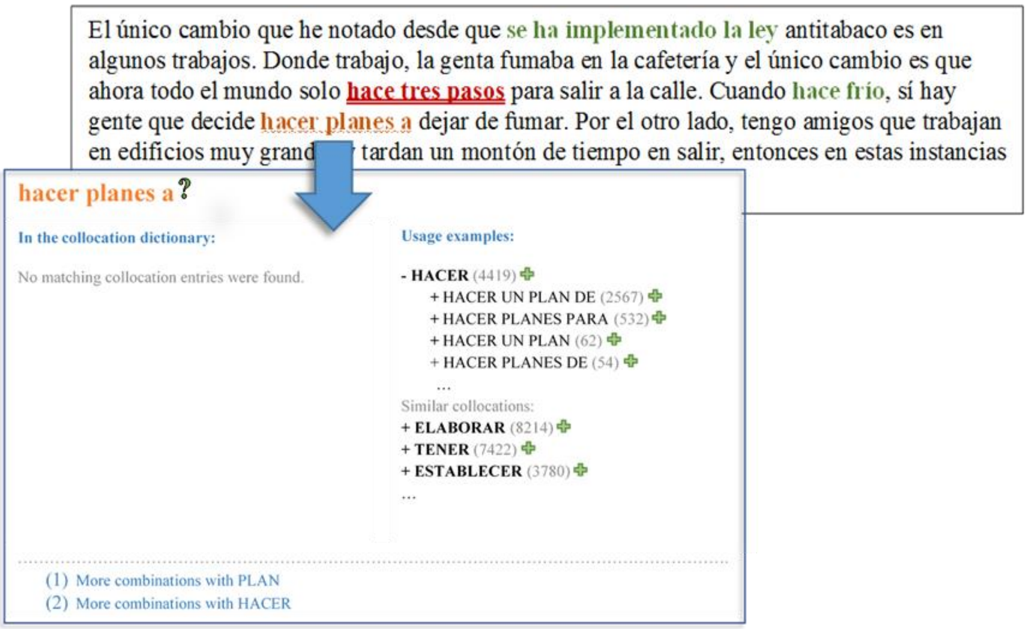


Figure 74 Collocation verification process

Figure 74 shows the proposed verification process to be followed by the learning tool to decide which combinations to mark for revision in the learner text. Collocations with a lexical entry in the dictionary database are marked as correct, while those not registered neither in the dictionary nor in the corpus module are marked for revision. Collocations attested in the corpus can be considered as correct if the frequency or the association strength of the combination is above a given threshold, otherwise they are marked for revision. Note that through matching a combination produced by a learner against the usage patterns attested in the corpus, it is possible to detect grammatical collocation errors. As shown in Figure 73, combinations with lexical choice errors, such as the case of **hace* *tres pasos* lit. 'makes three steps' are marked differently (double line) from combinations with grammatical errors, such as **hacer planes a* lit. 'make plans to' (dotted line) where an incorrect preposition is used.

As it is suggested above, through clicking on a collocation highlighted in their text, users are offered direct access to relevant information as part of the collocation verification feedback, see Figure 75. This is done through displaying the results of the collocation search corresponding to the combination selected by the user. Consequently, the information provided to the user relative to each individual combination is structured in the same way as it was explained in the previous section, taking into account whether the collocation introduced by the learner has a corresponding entry in the dictionary database, whether occurrences of it can be found in the corpus, and whether it reaches the threshold established for native-like combinations.



El único cambio que he notado desde que se ha implementado la ley antitabaco es en algunos trabajos. Donde trabajo, la genta fumaba en la cafetería y el único cambio es que ahora todo el mundo solo hace tres pasos para salir a la calle. Cuando hace frío, sí hay gente que decide hacer planes a dejar de fumar. Por el otro lado, tengo amigos que trabajan en edificios muy grand tardan un montón de tiempo en salir, entonces en estas instancias

hacer planes a ?

In the collocation dictionary:

No matching collocation entries were found.

Usage examples:

- HACER (4419)
  + HACER UN PLAN DE (2567)
  + HACER PLANES PARA (532)
  + HACER UN PLAN (62)
  + HACER PLANES DE (54)
  ...

Similar collocations:

+ ELABORAR (8214)
+ TENER (7422)
+ ESTABLECER (3780)
...

(1) More combinations with PLAN
(2) More combinations with HACER

Figure 75 Displaying collocation information in the “collocation verification” search scenario

As for further aid offered to users when it comes to correcting errors, two additional features were mentioned in 5.3.3.5. The first of these concerned implementing multi-step feedback, whereby users of the learning tool are first provided with the opportunity to contrast language data provided by the tool with their own production autonomously, and second, if they have difficulty in identifying the error, they can request additional help. This option can be especially relevant in the case of grammatical errors, which, as we have seen, appear to be less salient. In order to implement error specific feedback, the part of speech and morphological information used in error detection can be exploited to display an indication of the nature of the error. The second type of aid mentioned referred to the integration of the learning tool with external resources, in particular, dictionaries to be consulted by the user to find information on unknown vocabulary as well as on certain aspects that are not necessarily exemplified sufficiently in the corpus, such as irregular verb forms, use of prepositions, etc.

6.3 Modules supporting personalized collocation learning

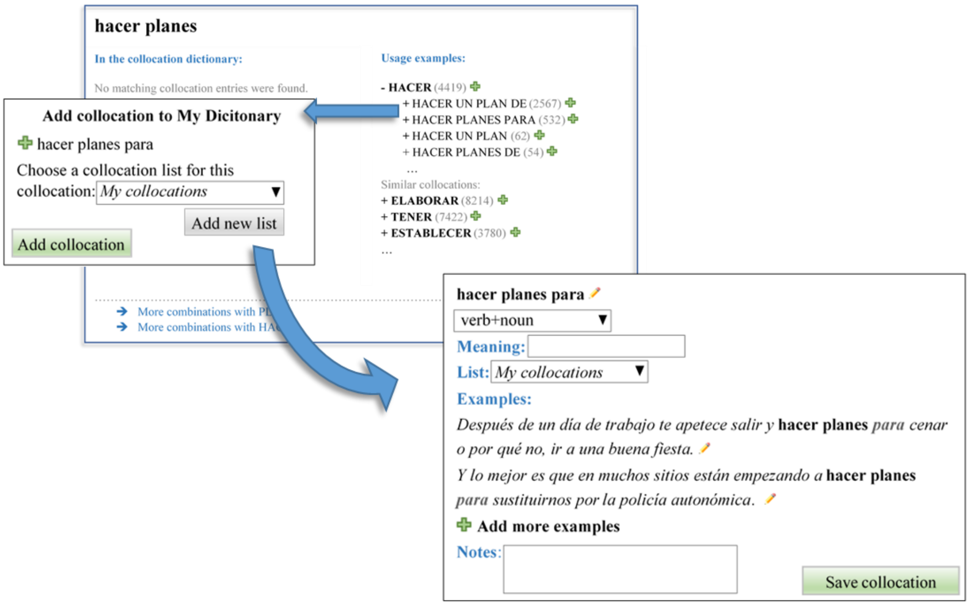
The previous section discussed the types of information offered by the collocation learning tool in different search scenarios, when used as a reference tool. We have seen that, in addition to queries resembling more traditional collocation dictionary look-ups, such as searching for a collocation or the list of collocates of a given lexical item, the tool can be used as a writing-aid, providing a more personalized learning experience. The aim of this section is to explore ways to exploit the lexical database underlying the dictionary module and complemented by corpus data to create an interactive learning environment, similarly to what is proposed by Verlinde et al. (2005). Accordingly, two additional components of the collocation learning tool are described briefly: a personal collocation dictionary and an activities module providing users with automatically generated collocation exercises, which allow practicing and memorizing word combinations.

6.3.1 Personal collocation dictionary

Woolard (2000, 43–44) argues that vocabulary notebooks constitute a fundamental learning tool, and that their only purpose should not be that of a place to note down new vocabulary. Instead, they should also be actively used to organize and revise lexical information, as well as to aid language production. This author also emphasizes that, in addition to single lexical items, vocabulary notebooks should also include collocations. Naturally, in an electronic learning environment, the paper vocabulary notebook can be substituted by a digital notebook, a personal space where the learner can collect and organize combinations encountered while using the interface. We have seen in 3.4.2.3.D that this idea is implemented in *FLAX* as the *cherry picking* functionality.

Similarly to *FLAX*, the learning tool described here incorporates a personal collocation dictionary component, to which users can add content while interacting with the tool. Collocations added to the users' personal dictionary can be organized into lists representing thematic groups. Importantly, as it is shown in Figure 76, combinatorial information is structured in categories similar to those used to include lexical data in the dictionary module, which allows better integration with other functionalities. When adding a collocation to the personal dictionary, a corresponding collocation entry is generated, which can be edited by the user, e.g. through changing or adding examples, translation equivalents or notes.

Figure 76 shows an example for adding the collocation *hacer planes* to the personal dictionary from the corpus module. While, saving a collocation from the dictionary module consists of copying an existing lexical entry and allowing users to adopt it to their own needs, adding an expression to the personal dictionary from the corpus module implies generating a new collocation entry. Once the user assigns the combination to a collocation list, the collocation entry is created through automatically determining the values of some of the data categories, such as *syntactic pattern* and *examples*. The user can optionally add a gloss or a translation equivalent to describe the meaning of the collocation, as well as modify existing data.



hacer planes

In the collocation dictionary:

No matching collocation entries were found.

Usage examples:

- HACER (4419)
  - + HACER UN PLAN DE (2567)
  - + HACER PLANES PARA (532)
  - + HACER UN PLAN (62)
  - + HACER PLANES DE (54)
  - ...

Similar collocations:

+ ELABORAR (8214)
+ TENER (7422)
+ ESTABLECER (3780)
...

More combinations with PL
More combinations with HA

Add collocation to My Dicitonary

hacer planes para

Choose a collocation list for this collocation: *My collocations*

hacer planes para

verb+noun

Meaning:

List: *My collocations*

Examples:

Después de un día de trabajo te apetece salir y **hacer planes** *para cenar o por qué no, ir a una buena fiesta.*

Y lo mejor es que en muchos sitios están empezando a **hacer planes** *para sustituirnos por la policía autonómica.*

Add more examples

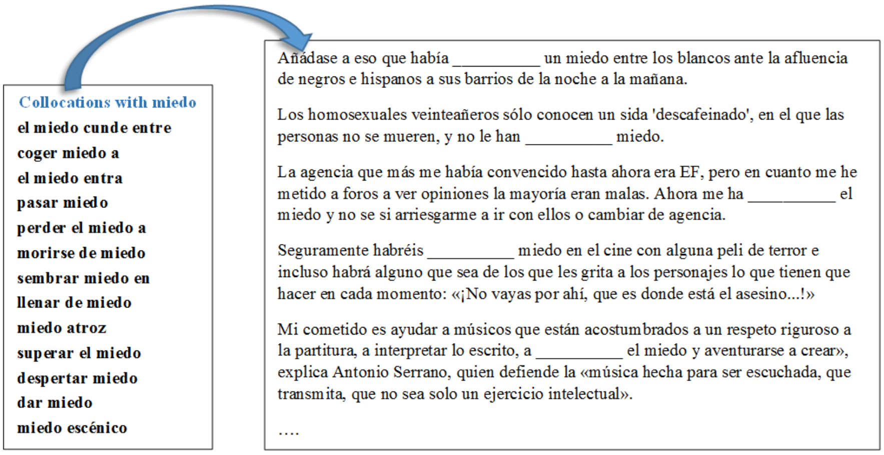
Notes:

Figure 76 Adding a collocation to the user's personal dictionary

In order to more fully integrate users' personal dictionary with the learning tool, it is possible to display its content together with that of the dictionary module in search results. At the same time, the content of the personal dictionary can also be exploited to generate collocation activities. Furthermore, users can be provided with the option of making their thematic collocation lists public, accessing, studying or collaboratively editing lists created by other users. Finally, entries generated by users can be used to enrich the content of the dictionary module through incorporating crowdsourced lexical information.

6.3.2 Learning activities

In addition to the creation of a personal collocation dictionary, users of the learning tool can have option of generating learning activities through selecting a set of collocations they desire to practice or memorize. In 3.4.2.3.D I described the activities included in *FLAX*, which are generated by the interface on the basis of corpus data. In the case of the learning tool *ALFALEX*[44], aimed at learners of French, Verlinde et al. (2005, 24–26) suggested exploiting the lexical database of an electronic dictionary as well as corpus data for automatically generating exercises involving collocations. According to this proposal, as a first step, relevant data, i.e. a list of target collocations, is chosen from the dictionary, then sentences containing these combinations are selected from the corpus and displayed in the form of a gap-fill exercise, and, finally, the system provides feedback to users' answers.



Collocations with miedo

el miedo cunde entre
coger miedo a
el miedo entra
pasar miedo
perder el miedo a
morirse de miedo
sembrar miedo en
llenar de miedo
miedo atroz
superar el miedo
despertar miedo
dar miedo
miedo escénico

Añádase a eso que había __________ un miedo entre los blancos ante la afluencia de negros e hispanos a sus barrios de la noche a la mañana.

Los homosexuales veinteañeros sólo conocen un sida 'descafeinado', en el que las personas no se mueren, y no le han __________ miedo.

La agencia que más me había convencido hasta ahora era EF, pero en cuanto me he metido a foros a ver opiniones la mayoría eran malas. Ahora me ha __________ el miedo y no se si arriesgarme a ir con ellos o cambiar de agencia.

Seguramente habréis __________ miedo en el cine con alguna peli de terror e incluso habrá alguno que sea de los que les grita a los personajes lo que tienen que hacer en cada momento: «¡No vayas por ahí, que es donde está el asesino...!»

Mi cometido es ayudar a músicos que están acostumbrados a un respeto riguroso a la partitura, a interpretar lo escrito, a __________ el miedo y aventurarse a crear», explica Antonio Serrano, quien defiende la «música hecha para ser escuchada, que transmita, que no sea solo un ejercicio intelectual».

….

Figure 77 Example for the generation of gap-fill exercises using corpus examples

In the case of the learning tool described here, target collocations for generating exercises can be selected from three sources: the dictionary module, the corpus module and the user's personal dictionary. Accordingly, users can perform searches for given lexical items and select the combinations and/or usage patterns they aim to practice from the results provided by the dictionary and corpus modules. At the same time, they can also opt for selecting a number of collocations from their personal dictionary, or the combinations included in a thematic collocation list to constitute the target items in a collocation activity. Figure 77 shows an example for a gap-fill exercise created from

[44] http://www.kuleuven.be/alfalex/index.php?id=&ng=0

corpus examples in order to practice the target combinations included in a user generated list containing collocations with the noun *miedo* ‘fear’.

6.4 Summary

The aim of this chapter was to propose a design for an online collocation learning tool through taking into account the findings of empirical studies presented previously in this thesis. The primary function of the interface described is that of a reference tool, which allows users to obtain combinatorial information from collocation dictionary and corpus simultaneously. With the aim of providing more flexible access to lexical data, it was proposed that, in the lexical database constituting the dictionary module, collocations should be classified using a comprehensive semantic typology, independent from syntactic classification, in addition to the more traditional grouping used generally in collocation dictionaries.

In addition, it was suggested that users should be able to query the reference tool via a single all-purpose search option which allows both single word and collocation searches. Single word searches provide access to collocation dictionary-type lexical entries, containing combinations in which the target item constitutes either the base or the collocate. Collocation searches retrieve the collocation entry corresponding to the search string, as well as provide opportunities to browse combinations related to it. The all-purpose search field also allows the user to introduce a text in order to exploit the interface as a writing aid tool focusing on collocations.

As for the description of collocations, it was noted that sufficient emphasis should be placed not only on lexical combinatory restrictions, but also on the grammatical features of the expressions in question. In the learner corpus study presented previously, it was found that many learner errors concern issues such as the use of articles, prepositions and government pattern specific to a combination. At the same time, the experiment dealing with autonomous error correction provided evidence that grammatical features are less salient in authentic language data. Consequently, it was suggested that morphosyntactic and syntactic information relevant for the correct formulation of a given combination should be systematically described in collocation entries, while corpus data presented to learners should also be meaningfully organized through emphasizing the most common usage patterns.

The final section of the chapter explored ways of exploiting lexical data underlying the learning tool in order to implement functionalities providing a more personalized learning experience. The two components described briefly correspond to a personal collocation dictionary in which the user can collect and organize word combinations, and an activities module which serves to generate collocation exercises based on a set of combinations selected by the learner.

Chapter 7. Conclusions

7.1 Introduction

The present thesis aimed to contribute to the body of research on foreign language learners' knowledge and use of multiword expressions. More precisely, it focused on the case of SFL learners and the subtype of multiword expressions constituted by collocations. In this concluding chapter, I take stock of the main outcomes of the work presented in the thesis, acknowledge its limitations, and consider possible directions for further research.

7.2 Main findings

Research described in this thesis was guided by three main aims. The first of them concerned the study of SFL learners' collocation use through comparing it to that of native speakers and examining collocation errors. The second aim was to examine the design and functionalities of existing learning tools that can support collocation learning, as well as to study these tools from a usability point of view. Finally, the third aim consisted of using the findings of these studies to propose a design for an online collocation learning tool aimed at learners of Spanish. The following sections consider the outcomes related to each of these.

7.2.1 SFL learners' use of collocations

The first question addressed in this thesis concerned the collocation production of learners of Spanish. When it comes to comparing learners' collocation use to that of native speakers, the results of the learner corpus study presented in Chapter 4 corroborate observations made in previous studies, which revealed that learners' use of collocations can be described in relation to a series of different phenomena, instead of simply in terms of the general underuse of the type of expressions in question. Firstly, when taking into account all restricted lexical combinations, regardless of their syntactic pattern, data showed that native and non-native speakers used a similar amount of collocations. Secondly, as it was expected, learner essays were found to contain a smaller repertoire of collocation lemmas, what is more, the difference in lexical diversity was found to be more accentuated in the case of the restricted element, the collocate. Thirdly, data concerning the use of frequent collocates, especially combinations containing high frequency verbs, appears to lend support to claims made in the literature regarding learners' tendency to rely on, and heavily overuse a small number of items, while, fourthly, the under- and overuse patterns observed in the case of different collocation types, particularly VERB+NOUN$_{\text{COMP}}$ and NOUN+MODIFIER combinations, raise the question concerning whether and to what extent the prominence of certain L1 structures affects the use of L2 expressions. To this end, it should be recalled that while EFL learners have been found to underuse VERB+NOUN$_{\text{COMP}}$ combinations (Altenberg and Granger 2001; Howarth 1996; Laufer and Waldman 2011), data presented here concerning L1 English learners of Spanish shows a significant overuse of this combination type as compared to native Spanish speakers.

When assessing the correctness of SFL learners' collocation use, nearly one fourth of the total number of collocations identified in learner essays were judged erroneous. Results obtained from the detailed analysis of collocation errors revealed that the majority of lexical errors affected the collocate, thus confirming the hypothesis implied by the notion of collocation adopted in this thesis, i.e. that the choice of the lexically restricted element of the combination poses particular difficulty to language learners. Furthermore, results concerning the source of collocation errors were in line with previous studies showing that most lexical errors likely resulted from transfer (e.g. Laufer and Waldman 2011; Martelli 2006; Nesselhauf 2005). Finally, the error analysis, making use of a detailed error typology, also highlighted collocation error types that have not received particular attention, such as grammatical errors, lexical errors affecting the base and lexical errors involving the use of target language non-words. The first of these was found to be an especially prominent category, and, therefore, it was suggested that learning resources should put more emphasis on the syntactic and morphosyntactic features of collocations, as opposed to current focus where these expressions are considered to be problematic mainly from the perspective of lexical selection.

7.2.2 Collocation learning resources and their use

The second question addressed by this thesis aimed at exploring collocation learning resources available to language learners, with a specific focus on the characteristics and use of collocation dictionaries, the potential use of language corpora, as well as the design and functionalities of online corpus tools tailored to learners' needs. A number of collocation dictionaries and online tools aimed at EFL and SFL learners were described in Chapter 3 (see 3.4.2).

While we have seen that collocation dictionaries aim to provide learners with a comprehensive list of combinations – naturally, subject to space restrictions –, it was also observed that the detail of description regarding the meaning of combinations as well as their syntactic and morphosyntactic features varies across dictionaries. Lexicographical works also differ in whether they lemmatize collocations under the base – generally considered to be the strategy best aiding language production –, the collocate or both. It was also noted that electronic versions of commercial collocation dictionaries do not make use of the full potential of the electronic medium when it comes to enhancing dictionary access. As for corpus-based learning resources with the function of reference tools, it was observed that they differ in the available search options, the type of feedback as well as the detail and amount of information provided. As a consequence, each of these tools can be described as satisfying slightly different user needs, and/or lacking functionalities to satisfy others.

The few studies testing learners' use of collocation dictionaries, reviewed in 3.4.2.1.E, revealed that the lack of explicit indication of meaning and general unfamiliarity with the specific dictionary type hinders successful use. The first part of Chapter 5 (see 5.2) described a usability experiment focusing on the online Spanish collocation dictionary, DiCE. As it was noted, the outcomes of this study not only serve to enhance the user interface of the given dictionary, but are also transferable to the design of other dictionary interfaces or reference tools. These more general findings include the observation that participants seemed to prefer the default search option in the dictionary, which is similar to traditional dictionary look-ups. On the one hand, this suggests that users' interactions with electronic dictionary interfaces are guided by their experience regarding how dictionaries are generally used. On the other hand, users' reluctance to explore and use more advanced alternative search options leads to the conclusion that it is more desirable for an electronic dictionary to include an all-purpose search option, which allows users to more easily discover the full potential of dictionary searches. The results of the experiment also called attention to the fact that dictionaries should avoid the extensive use of linguistic terminology users are unlikely to be familiar with, and highlighted the importance of instruction in dictionary use as part of the L2 curriculum.

Experimental studies concerning the use of corpora in production tasks reviewed in 3.4.2.2.B showed encouraging results regarding L2 learner's ability to make use of corpus data in both online production and error correction tasks. The second study described in Chapter 5 (see 5.3) aimed at testing to what extent SFL learners were able to correct different types of collocation errors extracted from the learner corpus with the help of concordances. Results of the experiment showed that participants were able to make use of concordance lines and correct good part of the target errors. It was also observed that participants were more successful in correcting lexical errors than grammatical errors, probably due to the relative lack of salience of the latter, while the use of full sentence concordances were found to result in more successful error correction attempts than that of n-gram concordance format. It was argued that the outcome of this experimental study constitutes evidence that SFL learners are able to interpret authentic language data, and contrast it with their language production, which is promising when it comes to evaluating the potential efficacy of an online corpus tool targeting collocation learning, such as the one proposed in Chapter 6.

7.2.3 Proposal for an online collocation learning tool

The third question addressed in the thesis concerned the design of an online collocation learning tool aimed at SFL learners. The proposal describing this tool is based on the premise that its design should be founded on a needs-driven approach, i.e. it should observe the results of empirical studies concerning language learners' collocation knowledge and use, as well as those investigating the usability and learning outcomes resulting from interactions with certain learning resources. The proposed learning tool has as its primary function that of a reference tool, and integrates a collocation dictionary and a corpus tool – the two resources complementing each other when it comes to providing combinatorial information –, while it also offers users the option of creating their personal collocation dictionary and generating learning activities.

In accordance with the results of the corpus study described in Chapter 4, which highlighted the prevalence of grammatical collocation errors, it was proposed that when presenting collocations, besides lexical combinatorics, emphasis should also be placed on offering syntactic and morphosyntactic information. The proposal also attempted to cater to other error types less commonly observed by learning resources, such as the use of L2 non-words and lexical errors in the base, through suggesting the implementation of incremental and fuzzy search and the inclusion of direct access to different types of alternative combinations from the collocation entry. Besides, in order to provide a more dynamic access allowing users to explore the full range of collocations expressing a given meaning, I proposed the implementation of independent semantic and syntactic typologies in the dictionary database. According to my expectations, the novel access path resulting from the semantic typology may contribute to both enhancing learners' sense of restricted lexical selection and broadening their collocation repertoire, though emphasizing meaning over syntactic pattern, thus allowing to discover expressions whose pattern does not match that of L1 equivalents.

Since, as it was mentioned above, the usability experiment presented in 5.2 suggested that participants had difficulties in using multiple search options, the reference tool was proposed to contain a single all-purpose search field which allows introducing single lexical items as well as full collocations. In addition, given the results of the study concerning collocation error correction showed that grammatical features of combinations tend to lack salience, it was suggested that, in addition to overtly describing such information in collocation dictionary entries, data offered by the corpus tool should be organized in a way that usage patterns, represented by n-grams, are given more prominence. Finally, since the n-gram format was on occasions found to result in misinterpretations, full sentence concordances, providing the broader context of target expressions should be always available to the user.

7.3 Limitations and suggestions for future research

The present thesis has taken a rather broad perspective in approaching different aspects of the issue of collocations in SLA. Consequently, it is clear that it would be possible to pursue further, more in-depth, research relevant to each of the three main questions considered. The following sections highlight the limitations of the empirical studies described in the thesis and enumerate some issues that deserve further inquiry.

7.3.1 Studying collocation knowledge and use

The learner corpus study presented in Chapter 4 provided some insight into how SFL learners' collocation use relates to that of native speakers, as well as the nature of collocation errors. Some of the limitations of the study result from the amount of corpus data used. Clearly, through analyzing a larger corpus, future research can obtain more reliable data and more in-depth insights into aspects of the collocation use of learners. In particular, one of the aims of my study was to take a holistic approach to collocations, as opposed to previous research generally limited to studying particular types of combinations, in most cases, VERB+NOUN$_{\text{COMP}}$ and MODIFIER+NOUN collocations. Nevertheless, given the low number of occurrences of combinations corresponding to other syntactic patterns, it was not possible to draw conclusions regarding how learners and native speakers compare in their use, or concerning the observed error rate. Consequently, it would be necessary to study a larger dataset potentially containing a higher amount of occurrences.

While a larger corpus allows to obtain more occurrences of the expressions studied, the use of different types of datasets can provide the opportunity to examine other aspects of learner collocations. In this study, I compared the collocation use of SFL learners to that of native speakers, however, future research should also involve subcorpora corresponding to SFL learners with different proficiency levels, in order to observe developmental features and uncover more of the characteristics of learners' collocation knowledge. As it was mentioned in 3.3.2, to my best knowledge, relatively few studies have dealt with this issue (see Howarth 1998b; Kaszubski 2000; Laufer and Waldman 2011). When it comes to overuse and underuse patterns, it can be worth carrying out a cross-linguistic analysis in order to elaborate on one of the findings of my study. As it was mentioned above, the results of this study showed that L1 English SFL learners overuse VERB+NOUN$_{\text{COMP}}$ collocations, while the same type of combinations are generally reported to be underused by EFL learners. This suggests that contrasting the data from the CEDEL2 corpus with L1 Spanish EFL learners’ production might provide more insight into L1 specific issues. Furthermore, a study involving SFL learners with different L1 backgrounds might allow to uncover further usage patterns.

Regarding the methodological aspects of corpus annotation, it should be noted that, since the collocation typology used in the error analysis was derived from the corpus data itself, its application to further data sets could serve to validate its generalizability. The typology cannot only be used to describe errors found in corpora, but also to analyze collocation errors occurring in production tests. A related methodological issue concerns the identification of the source of errors in the error analysis. Since on the basis of corpus data it is not possible to reliably judge the extent of L1 influence, or test hypotheses concerning production strategies, further research in this regard should involve studies adopting different methodological approaches to obtain a better insight.

7.3.2 Exploring collocation learning resources and their use

The main limitations of the usability study of the DiCE interface are related to the lack of sufficient control over participants' reference skills, and the nature of the test items, some of which involved dictionary searches that did not represent real-life look-up situations. In evaluating participants' interactions with the dictionary, considerable individual differences were observed in participants' ability to use the search interface. The study aimed to correlate these differences with user profiles, however, in order to obtain more reliable data concerning usability, future tests should involve a more thorough evaluation of participants' reference and user skills, referring to both their dictionary use habits and general internet and computer skills. This would be especially useful when it comes to assessing the relevance of the alarmingly poor performance of some participants.

As for the nature of test items, it was explained that the usability experiment aimed at providing opportunities to manipulate each search option and most search filters found in the DiCE interface. Consequently, not all test items were necessarily formulated in a way or required retrieving information that would correspond to likely look-ups carried out by the user groups represented by participants, and this could have affected the outcomes of the experiment. In a future test, dictionary use should be assessed in a more realistic setup through using the dictionary log to explore actual user interactions, not triggered by a specific questionnaire, and through experiments involving production tasks or other learning activities that prompt participants to carry out different look-ups relevant in an L2 context. The dictionary log allows to examine in what ways DiCE is used in an unintrusive manner, while a more task oriented test format can be used not only to test this particular dictionary but also to compare the use of different combinatory dictionaries, as well as other types of resources.

As for the limitations of the experiment concerning error correction with the help of concordance data, it should be noted that the full sentence and n-gram concordance formats were compared on a rather small scale – only involving seven out of the total number of twenty questionnaire items –, while the experiment setup did not allow to determine how successfully SFL learners themselves can query a tool providing concordance data or correct their own errors. These are all aspects that can be elaborated on in future studies. In order to make a more rigorous comparison between the efficacy of full sentence and n-gram concordance formats, these should be contrasted in the case of a higher amount of test items. Furthermore, the efficacy of a combined feedback type, where users access full sentence concordances through shorter segments representing usage patterns, such as in the case of the learning tool proposed in Chapter 6, should also be tested.

SFL learners' autonomous use of corpus data as a tool aiding language production can be further observed in a study where participants are asked to retrieve concordance lines themselves, instead of being provided a set of corresponding examples as it was done in the study described in 5.2. In such a study, participants can either query a generic concordance tool or a corpus tool that is specifically tailored to their need – such as those described in 3.4.2.3 or the one proposed in Chapter 6.

Note also that, since the aim of the study discussed in this thesis was to obtain data concerning the success of error correction in the case of specific types of errors, participants were asked to reformulate learner sentences coming from the CEDEL2 corpus. However, a study assessing concordance feedback from a more generic perspective, or evaluating the use of a specific tool should explore participants' performance in correcting their own production errors, as well as the use of corpus tools in a genuine language production task. This latter aspect is also interesting regarding the viability of a learning tool, in the sense that a scenario when learners can consult a concordance tool freely, whenever they believe it to be helpful, can demonstrate their interest in applying the tool to support their autonomous production. It was noted in this respect that Gaskell and Cobb (2004) observed that their subjects' concordance searches declined once they were not explicitly instructed to use this method.

7.3.3 Designing a collocation learning tool

Chapter 6 of this thesis described the design of a collocation learning tool which intended to take into account the user-needs outlined by the results of empirical studies. A task future work should deal with is that of verifying the effectiveness of the proposed features, such as the combined concordance feedback type, already mentioned in the previous section. In general, I believe that testing different features through the use of prototype versions should constitute an integral part of the process of building an actual tool.

One question of particular interest that was raised regarding the design of the dictionary module referred to the viability of the creation of a comprehensive semantic taxonomy of collocations that could be interpreted with sufficient ease by language learners. Approaching this issue requires an in-depth semantic analysis of combinations covering different semantic fields, as well as testing the typology with potential dictionary users. Furthermore, it is also necessary to establish whether the novel access path allowing to search collocations through their meaning does in fact result useful or entails the expected benefits.

7.4 Conclusion

The present thesis has examined SFL learners' needs when it comes to enhancing their collocation competence and use, with a view to designing a new collocation learning tool. The results of the learner corpus study described in Chapter 4 gave an insight into the problematic aspects of collocation use and provided information regarding the types of errors affecting learner collocations. At the same time, the usability test involving the DiCE interface and the experiment assessing learners' ability to correct collocation errors with the aid of concordance lines, described in Chapter 5, served to evaluate SFL learners' skills when it comes to interacting with different learning tools in an autonomous manner. As demonstrated in Chapter 6, the results of these studies can indeed prove beneficial when it comes to designing the features of a new learning tool.

References

Aisenstadt, Ester. 1979. Collocability restrictions in dictionaries. *ITL Review of Applied Linguistics* 45-46: 71–74.

———. 1981. Restricted collocations in English lexicology and lexicography. *ITL Review of Applied Linguistics* 53: 53–61.

Allerton, David. J. 1982. *Valency and the English verb*. London: Academic Press.

———. 1984. Three (or four) levels of word cooccurence restriction. *Lingua* 63 (1): 17–40.

Almela, Moisés. 2011. Improving corpus-driven methods of semantic analysis: A case study of the collocational profile of incidence. *English Studies* 92 (1): 84–99.

Alonso Ramos, Margarita. 1993. Las funciones léxicas en el modelo lexicográfico de I. Mel'čuk. (Doctoral dissertation). Universidad Nacional de Educación a Distancia, Madrid.

———. 1994. Hacia una definición del concepto de colocación: De J. R. Firth a I. A. Mel'čuk. *Revista de Lexicografía* 1: 9–28.

———. 1998. Étude sémantico-syntaxique des constructions à verbe support. (Doctoral dissertation). Université de Montréal, Montréal.

———. 2002. Colocaciones y contorno de la definición lexicográfica. *Lingüística Española Actual* 24 (1): 63–96.

———. 2003. Hacia un diccionario de colocaciones del español y su codificación. In Martí Antonín, M. Antònia, Fernández Montraveta, Ana and Vázquez García, Glòria (eds.), *Lexicografía computacional y semántica*, pp. 11–34. Barcelona: Edicions Universitat de Barcelona.

———. 2004. *Diccionario de Colocaciones del Español.* Retrieved from: www.dicesp.com.

———. 2006. Entón, ¿é unha colocación ou non?. *Cadernos de Fraseoloxia Galega* 8: 29–43.

———. 2008. Papel de los diccionarios de colocaciones en la enseñanza de español como L2. In Bernal, Elisenda and DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress*, pp. 1215–30. Barcelona: IULA, Documenta Universitaria. Retrieved from: http://www.euralex.org/elx_proceedings/Euralex2008/000_Euralex_2008_PRINTED_BOOK_TOC_Foreword_PlenaryLectures_Abstracts.pdf

———. 2009. Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario? In Cantos Gómez, Pascual and Sánchez Pérez, Aquilino (eds.), *A survey of corpus-based research. Panorama de investigaciones basadas en corpus*, pp. 1191–1207. Murcia: AELINCO.

———. 2011. Sobre los usos figurados: ¿extensiones de una única definición? In Escandell Vidal, M. Victoria, Leonetti, Manuel and Sánchez López, Cristina (eds.), *60 Problemas de Gramática Dedicados a Ignacio Bosque*, pp. 340–346. Madrid: Akal.

———. 2012. Naturaleza semántica de las colocaciones verbales. In Apresjan, Jurij, Boguslavsky, Igor, L'Homme, Marie-Claude, Iomdin, Leonid, Milićević , Jasmina, Polguère, Alain and Wanner, Leo (eds.), *Meaning, texts and other exciting things: Festschrift in honour of Igor Mel'čuk*, pp. 115–136. Moscow: Jazyki slavjanskoj kultury.

Alonso Ramos, Margarita, García Salido, Marcos and Vincze, Orsolya. 2014. Towards a collocation writing assistant for learners of Spanish. In Faaß, Gertrud and Ruppenhofer, Josef (eds.), *Workshop Proceedings of the 12th Edition of the Konvens Conference*, pp. 77–88. Hildesheim: Universitätsverlag Hildesheim. Retrieved from: http://www.uni-hildesheim.de/konvens2014/data/konvens2014-workshop-proceedings.pdf

Alonso Ramos, Margarita and Tutin, Agnès. 1996. A classification and description of lexical functions for the analysis of their combinations. In Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, pp. 147–278. Amsterdam/Philadelphia: John Benjamins.

Altenberg, Bengt and Granger, Sylviane. 2001. The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics* 22 (2): 173–194.

Amosova, Natalija. N. 1969. *Osnovui Anglijskoy Frazeologii*. Leningrad: University Press.

Ruiz Martínez, Ana María. 2006. REDES. Diccionario combinatorio del español contemporáneo como herramienta para la enseñanza del léxico en L2. *linRed: Lingüística En La Red* 4. Retreived from: http://www.linred.es/articulos_pdf/LR_articulo_21092006.pdf

Asmussen, Jörg. 2013. Combined products: Dictionary and corpus. In Gouws, Rufus H. Heid, Ulrich, Schweickard, Wolfgang and Wiegand, Herbert Ernst (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary Volume: Recent developments with focus on electronic and computational lexicography*, pp. 1081–90. Berlin/Boston: De Gruyter Mouton.

Bahns, Jens and Eldaw, Moira. 1993. Should we teach EFL students collocations? *System* 21 (1): 101–114.

Bannard, Colin and Lieven, Elena. 2012. Formulaic language in L1 acquisition. *Annual Review of Applied Linguistics* 32: 3–16.

Barrios Rodríguez, María A. 2007. Diccionarios combinatorios del español: diferencias y semejanzas entre *Redes* y *Práctico*. *RedELE. Revista Electrónica de Didáctica / Español Lengua Extranjera* 11. Retreived from: http://www.mecd.gob.es/dctm/redele/MaterialRedEle/Revista/2007_11/2007_redELE_11_01Barrios.pdf?documentId=0901e72b80df2cb7

Benson, Morton. 1989. The structure of the collocational dictionary. *International Journal of Lexicography* 2 (1): 1–14.

———. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography* 3 (1): 23–34.

Benson, Morton, Benson, Evelyn and Ilson, Robert. 1986a. *BBI Combinatory Dictionary of English*. Amsterdam / Philadelphia: John Benjamins.

———. 1986b. *The BBI Combinatory Dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.

———. 1986c. *Lexicographic description of English*. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan and Finegan, Edward. 1999. *Longman grammar of spoken and written English*. Harlow, UK: Longman.

Biskup, Danuta. 1992. L1 influence on learners' renderings of English collocations. A Polish/German empirical study. In Arraud, Pierre J. L. and Bejoint, Henri (eds.), *Vocabulary and applied linguistics*, pp. 85–93. London: Macmillan.

Boers, Frank and Lindstromberg, Seth. 2012. Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics* 32: 83–110.

Bonk, William J. 2000. *Testing ESL Learners' knowledge of collocations*. (Research report). University of Hawaii at Manoa, Honolulu.

Bosque, Ignacio. 2001. Sobre el concepto de 'colocación' y sus límites. *Lingüística Española Actual* 23 (1): 9–40.

———. 2004a. Combinatoria y significación. Algunas reflexiones. In Bosque, Ignaica (coord.) *Redes. Diccionario combinatorio del español contemporánero*, pp. LXXVII–CLXXIV. Madrid: S.M.

———. (coord.). 2004b. *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: S.M.

———. (coord.). 2006. *Diccionario combinatorio práctico del español contemporáneo: las palabras en su contexto*. Madrid: S.M.

———. 2011. Deducing collocations. In Boguslavsky, Igor and Wanner, Leo (eds.), *Proceedings of the 5th International Conference on Meaning-Text Theory*, pp. vi–xxiii. Retrieved from: http://meaningtext.net/mtt2011/proceedings/

Buendía Castro, Miriam and Faber, Pamela. 2014. Collocation dictionaries: A comparative analysis. *MonTi: Monografías de Traducción e Interpretación* 6: 203–235. Retrieved from: http://www.e-revistes.uji.es/index.php/monti/article/view/1673/1458

Bullon, Stephen. 2004. *Longman Dictionary of Contemporary English. 4th edition*. Harlow, Essex: Longman.

Carlini, Roberto, Codina-Filba, Joan and Wanner, Leo. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. In Volodina, Elena, Borin, Lars and Pilán, Ildikó (eds.), *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning at SLTC 2014, Uppsala University*, pp. 1–12. Uppsala: Uppsala University. Retrieved from: http://www.aclweb.org/anthology/W14-3501

Castillo Carballo, María Auxiliadora. 1998. El término ‘colocación’ en la lingüística actual. *Lingüística Española Actual* 20 (1): 41–54.

Álvarez Cavanillas, José Luis. 2008. Algunas aplicaciones del enfoque léxico al aula de ELE. (Master’s thesis). Universidad de Barcelona, Barcelona. Retrieved from: http://www.mecd.gob.es/redele/Biblioteca-Virtual/2008/memoriaMaster/1-Semestre/ALVAREZ_C.html

Chambers, Angela and O’Sullivan, Íde. 2004. Corpus consultation and advanced learners’ writing skills in French. *ReCALL* 16 (01): 158–172.

Chan, Tun-pei, and Liou, Hsien-Chin. 2005. Effects of web-based concordancing instruction on EFL students’ learning of verb–noun collocations. *Computer Assisted Language Learning* 18 (3): 231–251.

Chang, Yu-Chia, Chang, Jason S., Chen, Hao-Jan and Liou, Hsien-Chin. 2008. An automatic collocation writing assistant for taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning* 21 (3): 283–299.

Cobb, Tom. 1997. Is there any measurable learning from hands-on concordancing? *System* 25 (3): 301–315.

Conklin, Kathy and Schmitt, Norbert. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32: 45–61.

Conzett, Jane. 2000. Integrating collocation into a reading and writing course. In Lewis, Michael (ed.) *Teaching collocation: Further developments in the Lexical Approach*, pp. 70–87. Hove: Language Teaching Publications.

Corpas Pastor, Gloria. 1996. *Manual de Fraseología Española*. Madrid: Gredos.

———. 2001. Apuntes para el estudio de la colocación. *Lingüística Española Actual*, 23 (1): 41–56.

———. 2015. Register-specific collocational constructions in English and Spanish: A usage-based approach. *Journal of Social Sciences*. Retrieved from: http://thescipub.com/PDF/ofsp.9994.pdf

Coseriu, Eugeniu. 1962. *Sistema, norma y habla*. *Teoría del lenguaje y lingüística general*. Madrid: Gredos.

———. 1977. *Principios de Semántica Estructural*. *Principios de Semántica Estructural*. Madrid: Gredos.

Cowie, Anthony. P. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2 (3): 223–235.

———. 1988. Stable and creative aspects of vocabulary. In Carter, Ronald and McCarthy, Michael (eds.), *Vocabulary and language teaching*, pp. 126–139. London: Longman.

———. 1991. Multiword units in newspaper language. In Granger, Sylviane (ed.), *Perspectives on the English lexicon: A tribute to Jacques van Roey*, pp. 101–116. Louvain-la-Neuve: Cahiers de l'Institut de Linguistique de Louvain.

———. 1992. Multiword lexical units and communicative language teaching. In Arnaud, Pierre J. and Bejoint, Henri (eds.), *Vocabulary and Applied Linguistics*, pp. 1–12. London: Palgarve Macmillan.

———. 1994. Phraseology. In Asher, Ron E. (ed.), *The encyclopedia of language and linguistics*, pp. 3168–3171. Oxford: Oxford University Press.

———. 1998a. Phraseological dictionaries: Some east-west comparisons. In Cowie, Anthony P. (ed.), *Phraseology: Theory, Analysis and Applications*, pp. 209–228. Oxford: Pergamon.

———. 1998b. A.S. Hornby: A centenary tribute. In Fontenelle, Thierry, Hiligsmann, Philippe, Michiels, Archibald, Moulin, André, Theissen, Siegfried (eds.), *Proceedings of the 8th EURALEX International Congress: Part 1*, pp. 3-16. Retrieved from: http://www.euralex.org/elx_proceedings/Euralex1998_1/Anthony COWIE-A. S. Hornby_ a Centenary Tribute.pdf

Crowther, Johnathan, Dignen, Sheila and Lea, Diana. 2002. *Oxford Collocations Dictionary for students of English*. 1st edition. Oxford: Oxford University Press.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Dahlmeier, Daniel and Tou Ng, Hwee. 2011. Correcting semantic collocation errors with L1-Induced paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011*, Edinburgh, United Kingdom, pp. 101-117. Retrieved from: https://www.comp.nus.edu.sg/~nght/pubs/emnlp11_sem.pdf

de Hoyos Puente, José Carlos and Villar Díaz, María Belén. 2006. Utilidad del diccionario contextual en la enseñanza del español a estudiantes franceses. In Bruña Cuevas, Manuel, Caballos Bejano, María de Gracia, Illanes Ortega, Inmaculada, Ramírez Gómez, Carmen and Raventós Barangé, Anna (eds.) *La cultura del otro: español en*

Francia, francés en España/La culture d l'autre: espagnol en France, français en Espagne, pp. 1033–1043. Sevilla: Universidad de Sevilla. Retrieved from: http://dialnet.unirioja.es/servlet/libro?codigo=502433

de Schryver, Gilles-Maurice. 2003. Lexicographers' dreams in the electronic dictionary age. *International Journal of Lexicography* 16 (2): 143-199.

de Schryver, Gilles-Maurice and Joffe, David. 2004. On how electronic dictionaries are really used. In Williams, Geoffrey and Vessier, Sandra (eds.), *Proceedings of the 11th EURALEX International Congress*, pp. 187–196. Lorient: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. Retrieved from: http://www.euralex.org/proceedings-toc/euralex_2004/

Durrant, Philip. 2008. High frequency collocations and second language learning. (Doctoral dissertation). University of Nottingham, Nottingham. Retrieved from: http://eprints.nottingham.ac.uk/10622/1/final_thesis.pdf.

Durrant, Philip and Schmitt, Norbert. 2009. To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching* 47 (2): 157–177.

Dziemianko, Anna. 2010. Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. *International Journal of Lexicography* 23 (3): 257–273.

———. 2011. Does dictionary form really matter? In Akasy, Kaoru and Uchida, Satoru (eds.), *ASIALEX2011 Proceedings. Lexicography: Theoretical and practical perspectives*, pp. 92–101. Kyoto: Asian Association for Lexicography. Retrieved from: https://repozytorium.amu.edu.pl/jspui/bitstream/10593/5783/1/Dziemianko_2011_Does%20dictionary%20form%20really%20matter.pdf

———. 2014. On the presentation and placement of collocations in monolingual English learners' dictionaries: Insights into encoding and retention. *International Journal of Lexicography* 27 (3): 259–279.

Edmonds, Phil. (n.d.) *Just the Word*. Available at: http://www.just-the-word.com/

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18 (1): 91–126.

———. 2002. Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24 (2): 143–188.

Ellis, Nick C. 2012. Formulaic Language and Second Language Acquisition: Zipf and the phrasal Teddy Bear. *Annual Review of Applied Linguistics* 32: 17–44.

Ellis, Nick C, Simpson-Vlach, Rita and Maynard, Carson. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly* 42 (3): 375–396.

Ellis, Rod and Gary Barkhuizen. 2005. *Analyzing Learner Language*. Oxford: Oxford University Press.

Erman, Britt and Warren, Beatrice. 2000. The idiom principle and the open choice principle. *Text: An Interdisciplinary Journal of the Study of Discourse* 20 (1): 29–62.

Evert, S. 2005. The statistics of word co-occurrences: Word pairs and collocations. University of Stuttgart. (Doctoral dissertation). University of Stuttgart, Stuttgart. Retrieved from: http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/

———. 2009. Corpora and collocations. In Lüdeling, Anke and Kytö, Merja (eds.), *Corpus linguistics: An international handbook, Vol.2.*, pp. 1212–1248. Berlin/New York: Mouton de Gruyter.

Eyckmans, June. 2009. Toward an assessment of learners’ receptive and productive syntagmatic knowledge. In Barfield, Andy and Gyllstad, Henrik (eds.), *Researching collocations in another language*, pp. 139–152. New York: Palgarve Macmillan.

Eyckmans, June, Boers, Frank and Stengers, Hélène. 2007. Identifying chunks: Who can see the wood for the trees? *Language Forum* 33 (2): 85–100.

Farghal, Mohammed and Obiedat, Hussein. 1995. Collocations: A neglected variable in EFL. *International Review of Applied Linguistics* 33 (4): 315–331.

Fernández Lázaro, Gisele. 2014. Enseñanza-aprendizaje de las colocaciones en nivel inicial (A1-A2). *MarcoELE: Revista de didáctica de español como lengua extranjera* 19. Retrieved from: http://marcoele.com/descargas/19/fernandez-colocaciones_a1_a2.pdf

Ferrando Aramo, Verónica. 2009. Materiales didácticos para la enseñanza-aprendizaje de las colocaciones: análisis y propuestas. (Master's thesis). Universidad Rovira i Virgili, Tarragona. Retrieved from: http://www.mecd.gob.es/dctm/redele/Material-RedEle/Biblioteca/2010_BV_11/2010_BV_11_1er_trimestre/2010_BV_11_11Ferrando.pdf?documentId=0901e72b80e1904a

———. 2012. Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE. (Doctoral dissertation). Universitat Rovira i Virgili, Tarragona. Retrieved from: http://www.tdx.cat/handle/10803/84025.

Ferraro, Gabriela, Nazar, Rogelio, Alonso Ramos, Margarita and Wanner, Leo. 2014. Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation*, 48 (1): 45-64.

Ferraro, Gabriela, Rogelio Nazar and Leo Wanner. 2011. Collocations : A challenge in computer assisted language learning. In Boguslavsky, Igor and Wanner, Leo (eds.), *Proceedings of the 5th International Conference on Meaning-Text Theory*, pp. 69-79. Retrieved from: http://meaningtext.net/mtt2011/proceedings/

Firth, John R. 1957. Modes of meaning. In Palmer, F.R. (ed.). *Papers in Linguistics 1934-1951*, pp. 190–215. Oxford: Oxford University Press.

———. 1968a. A synopsis of linguistic theory. 1930-55. In Palmer, F.R. (ed.). *Selected Papers of J. R. Firth 1952-1959*, pp. 168-–205. London: Longmans.

———. 1968b. Linguistic analysis as a study of meaning." In Palmer, F.R. (ed.). *Selected Papers of J. R. Firth 1952-1959*, pp. 12-26. London: Longmans.

Frankenberg-Garcia, Ana. 2012. Learners’ use of corpus examples. *International Journal of Lexicography* 25 (3): 273–296.

Franz, Alex and Brants, Thorsten. 2006. All our n-gram are belong to you. *Google Research Blog*. Retrieved from http://googleresearch.blogspot.com.es/2006/08/all-our-n-gram-are-belong-to-you.html.

García Platero, Juan Manuel. 2002. Aspectos semánticos de las colocaciones. *Lingüística Española Actual* 24 (1): 25–34.

García-Page, Mario. 2001. El adverbio colocacional. *Lingüística Española Actual* 23 (1): 110–114.

Gaskell, Delian and Cobb, Thomas. 2004. Can learners use concordance feedback for writing errors? *System* 32 (3): 301–319.

Gatbonton, Elizabeth and Segalowitz, Norman. 2005. Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review / La revue canadienne des langues vivantes* 61 (3): 325–353.

Gelbukh, Alexander and Kolesnikova, Olga. 2013. *Semantic analysis of verbal collocations with lexical functions*. Berlin/Heidelberg: Springer.

Geluso, Joe. 2013. Phraseology and frequency of occurrence on the web: Native speakers’ perceptions of Google-informed second language writing. *Computer Assisted Language Learning* 26 (2): 144–157.

Geluso, Joe and Yamaguchi, Atsumi. 2014. Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL* 26 (2014): 225–242.

Gitsaki, Christina. 1996. *The development of ESL collocational knowledge.* (Doctoral dissertation). University of Queensland, Brisbane. Retrieved from: http://espace.library.uq.edu.au/view/UQ:205374

Goossens, Vannina. 2005. Le noms de sentiment: Esquisse de typologie sémantique fondée sur les collocations verbales. *Lidil* 32: 103–121.

Goossens, Vannina, Grutschus, Anke, Kern, Beate and Melnikova, Elena. 2013. *Emolex, EmoConc, EmoLing. documentation méthodologique*. Retrieved from: http://emolex.u-grenoble3.fr/emoBase/doc/Methodo_complet_05_06_2013.pdf

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Cowie, Anthony P. (ed.), *Phraseology: Theory, analysis and applications*, pp. 145–160. Oxford: Oxford University Press.

———. 2002. "A bird's eye view of learner corpus research. In Granger, Sylviane Hung, Joseph and Petch-Tyson, Stephanie, *Computer learner corpora, second language acquisition and foreign language teaching*, pp. 3–33. Amsterdam/Philadelphia: John Benjamins.

Greenbaum, Sidney. 1970. *Verb-intensifier collocations in English. An experimental approach*. The Hague: Mouton de Gruyter.

———. 1974. Some verb-intensifier collocations in American and British English. *American Speech*, 49 (1-2): 79–89.

Gries, Stefan. 2013. 50-something years of work on collocations: What is or should be next …." *International Journal of Corpus Linguistics* 18 (1): 137–166.

Grimes, Joseph. 1990. Inverse lexical functions. In Steele, J. (ed.), *Meaning-Text Theory: Linguistics, lexicography and implications*, pp. 350–364. Ottawa: Ottawa University Press.

Gyllstad, Henrik. 2007. *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. (Doctoral dissertation). Lund University, Lund (Sweden). Retrieved from: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=599011&fileOId=2172422

Halliday, Michael A.K. 1961. Categories of the theory of grammar. *Word* 17: 241–292.

———. 1966. Lexis as a linguistic level. In Bazell, Charles E., Catford, C., Halliday, Michael A.K. and Robins, R.H. (eds.), *In Memory of J. R. Firth*, pp. 148–162. London: Longman.

Handl, Susanne. 2009. Towards collocational webs for presenting collocations in learners' dictionaries. In Barfield, Andy and Gullstad, Henrik (eds.), *Researching Collocations in Another Language*, pp. 69–85. New York: Palgarve Macmillan.

Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4 (2): 237–260.

Hausmann, Franz Josef. 1979. Un dictionnaire des collocations est-il possible? *Travaux de Linguistique et de Littérature* 17 (1): 187–195.

———. 1985. "Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Teorie des lexicographishen Beispiels. In Bergenholtz, Henning and Mugdon, Joachim (eds.), *Lexicographie und Grammatik*, pp. 118–129. Tübingen: Niemeyer.

———. 1989. Le dictionnaire de collocations. In Hausmann, Franz Josef, Reichmann, Oskar, Wiegand, Herbert Ernst and Zgusta, Ladislav (eds.), *Wörterbücher, Dictionaries, Dictionnaires, Vol. 1.*, pp. 1010–1019. Berlin: de Gruyter.

———. 1998. O diccionario de colocacións. Criterio de organización. In Ferro Ruibal, Xesús (ed.), *Actas do 1º Coloquio Galego de Fraseoloxía*, pp. 63–81. Vigo: Xunta de Galicia. Retrieved from: http://www.cirp.es/pub/docs/actas_coloquio_frase.pdf.

Heid, Ulrich. 2004. On the presentation of collocations in monolingual dictionaries. In Williams, Geoffrey and Vessier, Sandra (eds.), *Proceedings of the 11th EURALEX International Congress*, pp. 729–738. Lorient (France): Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. Retrieved from: http://www.euralex.org/proceedings-toc/euralex_2004/

———. 2011. Electronic dictionaries as tools: Towards an assessment of usability. In Fuentes-Oliveira, Pedro A. and Bergenholz, Henning (eds.), *E-lexicography: The Internet, digital initiatives and lexicography*, pp. 287–304. New York: Continuum.

Heid, Ulrich and Zimmermann, Jan Timo. 2012. Usability testing as a tool for e-dictionary design: Collocations as a case in point. In Fjeld, Ruth Vatvedt and Torjusen, Julie Matilde (eds.), *Proceedings of the 15th EURALEX International Congress*, pp. 661–671. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. Retrieved from: http://www.euralex.org/proceedings-toc/euralex_2012/

Henriksen, Birgit. 2013. Research on L2 learners' collocational competence and cevelopment – A progress report. In Bardel, Camilla, Lindquist, Christina and Laufer, Batia (eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, pp. 29–56. European Second Language Association. Retrieved from: http://www.eurosla.org/monographs/EM02/EM02home.php

Herbst, Thomas. 1996. What are collocations: Sandy beaches or false teeth? *English Studies* 77 (4): 379–393.

Higueras García, Marta. 2006. *Las colocaciones y su enseñanza en la clase de ELE*. Madrid: Arco Libros.

———. 2007. *Estudio de las colocaciones léxicas y su enseñanza en español como lengua extranjera*. Madrid: MEC, Secretaría General de Educación.

———. 2008. *El diccionario Práctico en la práctica docente del español como lengua extranjera*. Alicante: Biblioteca Virtual Miguel de Cervantes. Retrieved from: http://www.cervantesvirtual.com/obra/el-diccionario-prctico-en-la-prctica-docente-del-espaol-como-lengua-extranjera-0/

Hill, Jimmie. 2000. "Revising priorities: From grammatical failure to collocational success. In Lewis, Michael (ed.), *Teaching collocation: Further developments in the Lexical Approach*, pp. 47–69. Hove: Language Teaching Publications.

Hill, Jimmie, Lewis, Morgan and Lewis, Michael. 2000. Classroom strategies, activities and exercises. In Lewis, Michael (ed.), *Teaching collocation: Further developments in the Lexical Approach*, pp. 88–117. Hove: Language Teaching Publications.

Hoey, M. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.

Howarth, Peter Andrew. 1996. *Phraseology in English academic writing*. Tübingen: Max Niemeyer Verlag.

———. 1998a. The phraseology of learners' academic writing. In Cowie, Anthony P. (ed.) *Phraseology, theory, analysis and applications*, pp. 161–186. Oxford: Claredon Press.

———. 1998b. Phraseology and second language proficiency. *Applied Linguistics* 19 (1): 24–44.

Huang, Zeping. 2014. The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 Writing. *ReCALL* 26 (March 2014): 163–83.

Hussein, Fayez Riyad 1990. Collocations: The missing link in vocabulary acquisition amongst EFL learners. In Fisiak, Jacek (ed.), *Papers and studies in contrastive linguistics. The Polish-English contrastive project*, pp. 123–136. Poznan: Adam Mickiewicz University.

Instituto Cervantes. 2006. *Plan Curricular Del Instituto Cervantes*. Madrid: Biblioteca Nueva. Retrieved from: http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/default.htm.

———. 2014. Corpus de aprendices de español (CAES). Available at: http://galvan.usc.es/caes

Iordanskaja, Lidija, Kim, Myunghee and Polguère, Alain. 1996. Some procedural problems in the implementation of lexical functions for text generation. In Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, pp. 279–298. Amsterdam/Philadelphia: John Benjamins.

Johns, Susan and Sinclair, John. 1974. English lexical collocations. *Cahiers de Lexicologie* 24: 15–61.

Johns, Tim. 1986. Micro-Concord: A language learner's research tool. *System* 14 (2): 151–162.

———. Contexts: The background, development and trialing of a concordance-based CALL program. In Wichman, Anne, Fligelstone, Steven, McEnery, Tony and Knowles, Gerry (eds.), *Teaching and Language Corpora*, pp. 100–115. London: Longman, 1997.

———. 2000. Data-driven learning: The perpetual challenge. In Ketteman, Bernhard and Marko, Georg (eds.), *Teaching and learning by doing corpus analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000*, pp. 107–117. Amsterdam: Rodopi.

Jousse, Anne-Laure. 2010. *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. (Doctoral dissertation). Université de Montréal / Univerité Paris Diderot (Paris 7), Montréal. Retrieved from: http://olst.ling.umontreal.ca/pdf/TheseALJousse.pdf

Jousse, Anne-Laure, Marie-Claude L'Homme, Patrick Leroyer and Benoît Robichaud. 2011. "Presenting Collocates in a Dictionary of Computing and the Internet according to User Needs." In *Proceedings of the 5th International Conference on Meaning-Text Theory Barcelona, September 8 – 9*, edited by Igor Boguslavsky and Leo Wanner, 134–144. Barcelona.

Jousse, Anne-Laure and Polguère, Alain. 2005. *Le DiCo et sa version DiCouèbe: document descriptif et manuel d'utilisation*. Montréal: Université de Montréal. Retrieved from: http://olst.ling.umontreal.ca/dicouebe/DiCoDOC.pdf

Kaszubski, Przemyslaw. 2000. *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: A contrastive, corpus-based approach*. (Doctoral dissertation). Adam Mickiewicz Univeristy, Poznan. Retrieved from: http://www.staff.amu.edu.pl/~przemka/rsearch.html#PhD

Keshavarz, Mohammed Hossein and Salimi, Hossein. 2007. Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics* 17 (1): 81–92.

Kilgarriff, Adam. 2009. Corpora in the classroom without scaring the students. Presented at the *18th International Symposium on English Teaching, 13-15 November 2009,* *Taipei.* Retrieved from: http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel and Suchomel, Vít. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36.

Kilgarriff, Adam, Husák, Miloš, McAdam, Katy, Rundell, Michael and Rychlý, Pavel. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In Bernal, Elisenda and DeCesaris, Janet (eds.), *Proceedings of the 13th EURALEX International Congress*, pp. 425–432. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. Retrieved from: http://www.euralex.org/proceedings-toc/euralex_2008/

Kilgarriff, Adam, Marcowitz, Fredrik, Smith, Smith and Thomas, James. 2015. Corpora and language learning with the Sketch Engine and SKELL. *Revue Français de Linguistique Appliquée* 22 (1): 61-80.

Kilgarriff, Adam, Rychly, Pavel, Smrz, Pavel and Tugwell, David. 2004. The Sketch Engine. In Williams, Geoffrey and Vessier, Sandra (eds.), *Proceedings of the 11th EURALEX International Congress*, pp. 105-116. Lorient (France): Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. Retrieved from: http://www.euralex.org/proceedings-toc/euralex_2004/

Kilgarriff, Adam and Tugwell, David. 2001. Word Sketch: Extraction and display of significant collocations for lexicography for lexicography. In *Proceedings of the ACL Workshop on Collocation: Computational extraction, analysis and exploitation*, pp. 32-38. Toulouse: ACL. Retrieved from: http://www.kilgarriff.co.uk/Publications/2001-KilgTugwell-ACLcollos-Sketches.pdf

Kjellmer, Göran. 1984. Some thoughts on collocational distinctiveness. In Aarts, Jan and Meijs, Willem (eds.), *Corpus linguistics: Recent developments in the use of computer corpora in English language research*, pp. 163–171. Amsterdam: Rodopi.

———. 1991. A mint of phrases. In Aijmer, Karin and Altenberg, Bengt (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pp. 111–127. London: Longman.

———. 1994. *A dictionary of English collocations. Volume 1*. Oxford: Clarendon.

Kochmar, Ekaterina and Briscoe, Ted. 2015. Using learner data to improve error correction in adjective-noun combinations. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, June 4, 2015, Denver, Colorado*, pp. 233–242. Association for Computational Linguistics. Retrieved from: http://www.cs.rochester.edu/u/tetreaul/bea10proceedings.pdf

Koike, Kazumi. 2001. *Colocaciones léxicas en el español actual: Estudio formal y léxico-semántico*. Alicante/Tokio: Universidad de Alicante/Takushoku University.

———. 2002. Comportamientos semánticos en las colocaciones léxicas. *Lingüística Española Actual* 24 (1): 6–23.

Kolesnikova, Olga and Gelbukh, Alexander. 2010. Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish. In Sidorov, Grigori, Hernández Aguirre, Arturo and Garcia, Carlos Alberto (eds.), *Advances in Soft Computing : 9th Mexican International Conference on Artificial Intelligence, MICAI 2010, Pachuca, Mexico, November 8-13, 2010, Proceedings, Part2*, pp. 196–207. Berlin/Heidelberg: Springer.

Komuro, Yuri. 2009. Japanese learners' collocation dictionary retrieval performance. In Barfield, Andy and Gyllstad, Henrik (eds.), *Researching collocations in another language*, pp. 86–98. New York: Palgarve Macmillan.

Koya, Taeko. 2005. The acquisition of basic collocations by Japanese learners of English. (Doctoral dissertation). Waseda University, Tokyo. Retrieved from: http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/5285/3/Honbun-4160.pdf

Krashen, Stephen and Scarcella, Robin. 1978. On routines and patterns in language acquisition and performance. *Language Learning* 28 (2): 283–300.

Kuiper, Koenraad. 1996. *Smooth Talkers*. Mahwah, NJ: Lawrence Erlbaum.

Kuiper, Koenraad. 2004. Formulaic performance in conventionalised varieties of speech. In Schmitt, Norbert (ed.), *Formulaic sequences: Acquisition, processing and use*, pp. 37–54. Amsterdam/Philadelphia: John Benjamins.

L'Homme, Marie-Claude. 2009. *DiCoInfo: Le dictionnaire fondamental de l'informatique et de l'Internet. Dictionnaire élaboré par l'équipe ÉCLECTIC Observatoire de Linguistique Sens-Texte (OLST)*. Montréal: Université de Montréal. Retrieved from: http://olst.ling.umontreal.ca/dicoinfo/manuel-DiCoInfo.pdf

L'Homme, Marie-Claude and Laneville, Marie-Eve. 2009. *DiCoEnviro: Le dictionnaire fonamental de l'environnement*. *Dictionnaire élaboré par l'équipe ÉCLECTIC Observatoire de Linguistique Sens-Texte (OLST)*. Montréal: Université de Montréal. Retrieved from: http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf

L'Homme, Marie-Claude and Leroyer, Patrick. 2009. Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology* 15 (2): 258–283.

Landure, Corinne and Boulton, Alex. 2010. Corpus et autocorrection pour l'apprentissage ses langues. *ASp* 57: 11–30. Retrieved from: http://asp.revues.org/931

Laufer, Batia. 2010. The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography* 24 (1): 29–49.

Laufer, Batia and Waldman, Tina. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61 (2): 647–672.

Lew, Robert. 2012. How can we make electronic dictionaries more effective?" In Granger, Sylviane and Paquot, Magali. *Electronic lexicography,* pp. 343-362. Oxford: Oxford University Press.

———. 2013. Online dictionary skills. In Kosem, Istok, Kallas, Jelena, Gantar, Polona, Krek, Simon, Langemets, Margit and Tuulik, Maria (eds.), *Electronic lexicography in* *the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, pp. 16–31. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. Retrieved from: http://eki.ee/elex2013/proceedings/eLex2013_02_Lew.pdf

Lew, Robert and Mitton, Roger. 2011. Not the word I wanted? How online English learners' dictionaries deal with misspelled words. In Kosem, Iztok and Kosem, Karmen (eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10-12 November 2011*, pp. 165–174. Ljubljana: Trojina, Institute for Applied Slovene Studies. Retrieved from: http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-21.pdf

———. 2013. Online English learners' dictionaries and misspellings: One year on. *International Journal of Lexicography* 26 (2): 219–233.

Lew, Robert and Radłowska, Magdalena. 2010. Navigating dictionary space: The findability of English collocations in a general learner's dictionary (LDOCE4) and special-purpose dictionary of collocations (OCD). In Ciuk, Andrzej and Molek-Kozakowska, Katarzyna (eds.), *Exploring space: Spatial notions in cultural, literary and language studies. Volume 2: Space in language studies*, pp. 34–47. Newcastle upon Tyne: Cambridge Scholars Publishing.

Lewis, Michael. 1993. *The Lexical Approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

———. 1997. *Implementing the Lexical Approach: Putting theory into practice*. Hove: Heinle.

———. 2000a. Language in the Lexical Approach. In Lewis, Michael (ed.), *Teaching collocation: Further developments in the Lexical Approach*, pp. 126–154. Hove: Language Teaching Publications.

———. 2000b. Materials and resources for teaching collocation. In Lewis, Michael (ed.), *Teaching collocation: Further developments in the Lexical Approach*, pp. 186–204. Hove: Language Teaching Publications.

Lieven, Elena and Tomasello, Michael. 2008. Children's first language acquisition from a usage-based perspective. In Robinson, Peter and Ellis, Nick C. (eds.), *Handbook of cognitive linguistics and second language acquisition*, pp., 168–196. London: Routledge.

Liu, Anne Li-E, Wible, David and Tsao, Nai-Lung. 2009. automated suggestions for miscollocations." In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 47–50. Boulder, CO: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/W09-2107

Lorenz, Gunter R. 1999. *Adjective intensification - learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

Lozano, Cristobal. 2009. CEDEL2: Corpus escrito del español L2. In Bretones Callejas, Carmen M., Fernández Sánchez, José Francisco, Ibáñez Ibáñez, José Ramón, García Sánchez, María Elena, Cortés de los Ríos, Mª Enriqueta, Salaberri Ramiro, Sagrario, Cruz Martínez, Mª Soledad, Perdú Honeyman, Nobel and Cantizano Márquez, Blasina (eds.), *Applied linguistics now: Understanding language and mind / La lingüística aplicada hoy: Comprendiendo el lenguaje y la mente*, pp. 197–212. Almería: Universidad de Almería.

Lozano, Cristobal and Mendikoetxea, Amaya. 2013. Learner corpora and second language acquisition: The design and collection of CEDEL. In Díaz-Negrillo, Ana, Ballier, Nicolas and Thompson, Paul (eds.), *Automatic treatment and analysis of learner corpus data*, pp. 65–100. Amsterdam: John Benjamins.

Lux-Pogodalla, Veronika and Polguère, Alain. 2011. Construction of a French lexical network: Methodological issues. In *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011, August, 1-5, 2011, Ljubljana, Slovenia*, pp. 55-62. Retrieved from: http://alpage.inria.fr/~sagot/pub/WoLeR_2011_proceedings.pdf

Manning, Christopher D. and Schütze, Hindrich. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Martelli, Aurelia. 2006. A corpus based description of English lexical collocations used by Italian advanced learners. In Marello, Carla, Corino, Elisa and Onesti, Cristina (eds.), *Atti del XII Congresso Internazionale di Lessicografia*, pp. 1005–1011. Alessandria: Edizioni dell'Orso.

Martinez, Ron. 2013. A framework for the inclusion of multi-word expressions in ELT. *ELT Journal* 67: 184–198.

Mayor, Michael. (ed.). 2009. *Longman Dictionary of Contemporary English*. *5th edition.* 2009. Harlow: Pearson Longman.

Mayor, Michael. 2013. *Longman Collocations Dictionary and Thesaurus*. Pearson Longman.

McGee, Iain. 2012. Collocation dictionaries as inductive learning resources in data-driven learning - An analysis and evaluation. *International Journal of Lexicography* 25 (3): 319–361.

McIntosh, Colin, Francis, Ben and Poole, Richard. (eds.) 2009. *Oxford Collocations Dictionary for students of English*. *2nd edition*. Oxford: Oxford University Press.

Mel'čuk, Igor. 1981. Meaning-Text models: A recent trend in Soviet linguistics. *Annual Review of Anthropology* 10: 27–62.

———. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. In Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, pp. 37–102. Amsterdam/Philadelphia: John Benjamins.

———. 1998. Collocations and lexical functions. In Cowie, Anthony, P. (ed.), *Phraseology: Theory, analysis and applications*, pp. 23–52. Oxford: Claredon Press.

———. 2008. Phraséologie dans la langue et dans le dictionnaire. In Campà, Àngels and Baqué, Lorraine (eds.), *Repères & Applications (VI). XXIV Journées Pédagogiques sur l'Enseignement du Français en Espagne Barcelone, 3-5 Septembre 2007*, pp. 187–200. Barcelona: Institut de ciències de l'educació, Universitat Autònoma de Barcelona.

———. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3 (1): 31–56.

———. 2013. Tout ce que nous voulions savoir sur les phrasèmes, mais... *Cahiers de Lexicologie* 1: 129–149.

Mel'čuk, Igor. 2015. Lexical functions: Description of lexical relations in a lexicon. In *Semantics: From meaning to text*, 155–279. Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, Igor, Arbatchewsky-Jumarie, Nadia, Dagenais, Louise, Elnitsky, Léo, Iordanskaja, Lidija, Lefebvre, Maie-Noëlle and Mantha, Suzanne. 1988. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Arbatchewsky-Jumarie, Nadia, Elnitsky, Léo, Iordanskaja, Lidija and Lessard, Adèle. 1984. *dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Arbatchewsky-Jumarie, Nadia, Iordanskaja, Lidija and Mantha, Suzanne. 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Arbatchewsky-Jumarie, Nadia, Iordanskaja, Lidija, Mantha, Suzanne and Polguère, Alain. 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, Igor, Clas, André and Polguère, Alain. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Mel'čuk, Igor and Wanner, Leo. 1996. Lexical functions and lexical inheritance for emotion lexemes in German. In Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, pp. 209–278. Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, Igor and Žolkovskij, Alexander. 1970. Towards a functioning Meaning-Text model of language. *Linguistics* 57: 10–47.

Mendikoetxea, Amaya. 2014. Corpus-based research in second language Spanish. In Geeslin, Kimberly L. (ed.), *The handbook of Spanish second language acquisition*, pp. 11–29. Malden, MA: Wiley-Blackwell.

Milićević, Jasmina and Hamel, Marie-Josée. 2007. Un dictionnaire de reformulation pour apprenants avancés du français langue seconde. *Revue de l'Université Moncton* [Numéro hors-série], 145–67.

Milton, John. 2006. Resource-rich web-based feedback: Helping learners become independent writers. In Hyland, Ken and Hyland, Fiona (eds.), *Feedback in second language writing*, pp. 123–37. Cambridge: Cambridge University Press.

Milton, John and Cheng, Vivying S. Y.. 2010. A toolkit to assist L2 learners become independent writers. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing processes and authoring aids*, pp. 33–41. Los Angeles, CA: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/W10-0405

Mitchell, Terence Frederick. 1971. Linguistic 'goings-on': Collocations and other lexical matters arising on the syntagmatic record. *Archivum Linguisticum* 2: 36–69.

Mochizuki, Masamichi. 2002. Exploration of two aspects of vocabulary knowledge: paradigmatic and collocational. *Annual Review of English Language Education in Japan* 13: 121–29.

Moreno Jaén, María. 2007. Un modelo para la enseñanza de colocaciones de una lengua extranjera mediante herramientas digitales. In González Rey, Maribel (ed.), *Adquisición de las expresiones fijas. Metodología y recursos didácticos*, pp. 225–247. Fernelmont: E.M.E. & Intercommunications S.P.R.L.

———. 2008. Teaching collocations through DDL: Design, implementation and preliminary results of a corpus-based learning experience. In Frankenberg-Garcia, Ana, Rkibi, Tawfiq, Braga da Cruz, Maria do Rosário, Carvalho, Ricardo, Direito, Cristina and Santos-Rosa, Diogo. (eds.), *Proceedings of the 8th Teaching and Language Corpora Conference, Lisbon, Portugal, 03-06 July 2008,* pp. 231–238. Lisbon: Associação de Estudos e de Investigação Científica do ISLA. Retrieved from: http://anafrankenberg.synthasite.com/resources/TaLCLisbon2008Proceedings.pdf

———. 2009. *Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional*. (Doctoral dissertation). Universidad de Granada, Granada.

Moreno, Pol, Ferraro, Gabriela and Wanner, Leo. 2013. Can we determine the semantics of collocations without using semantics? In Kosem, Istok, Kallas, Jelena, Gantar, Polona, Krek, Simon, Langemets, Margit and Tuulik, Maria (eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, pp. 106–121. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. Retrieved from: http://eki.ee/elex2013/proceedings/eLex2013_08_Moreno+Ferraro+Wanner.pdf

Myles, Florence, Mitchell, Rosamond and Hooper, Janet. 1999. Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition* 21 (1): 49–80.

Nastase, Vivi, Sayyad-Shirabad, Jelber, Sokolova, Marina and Szpakowicz, Stan. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence,* pp. 781-786. Retrieved from: https://www.aaai.org/Papers/AAAI/2006/AAAI06-124.pdf

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nattinger, James R. and DeCarrico, Jaenette S. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Navajas Algaba, Aurora. 2006. *Las colocaciones en el aula de E/LE: Actividades para su explotación didáctica*. (Master’s thesis). Universidad Antonio de Nebrija, Madrid. Retrieved from: http://www.mecd.gob.es/dctm/redele/Material-RedEle/Biblioteca/2007_BV_08/2007_BV_08_19Navajas.pdf?documentId=0901e72b80e2d982

Nesselhauf, Nadja. 2004. What are collocations? In Allerton, David John, Nesselhauf, Nadja and Skandera, Paul (eds.), *Phraseological units: Basic concepts and their application*, pp. 1–21. Basel: Schwabe.

———. 2005. *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.

Nizonkiza, Déogratias. 2012. Quantifying controlled productive knowledge of collocations across proficiency and word frequency levels. *Studies in Second Language Learning and Teaching* 2 (1): 67–92.

Nuccorini, Stefania. 2003. Towards an ‘ideal’ dictionary of English collocations. In van Sterkenburg, Piet (ed.), *A practical guide to lexicography*, pp. 366–87. Amsterdam/Philadelphia: John Benjamins.

O’Sullivan, Íde and Chambers, Angela. 2006. “Learners’ writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing* 15 (1): 49–68.

Ogren, Philip V. 2006. Knowtator: A Protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the NAACL. Companion Volume*, 273–75. New York: Association of Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/N06-4006

Orol González, Ana and Margarita Alonso Ramos. 2013. A comparative study of collocations in a native corpus and a learner corpus of Spanish. *Procedia - Social and Behavioral Sciences* 96: 563–570. Retrieved from: http://www.sciencedirect.com/science/article/pii/S1877042813042031

Pacheco López, Myriam. 2003. El diccionario de colocaciones como herramienta en la enseñanza del español como lengua extranjera. *Linguax, Revista de Lenguas Aplicadas*, 1–29.

Palmer, Harold E. 1933. *Second interim report on English collocations*. Tokyo: Kaitakusha.

Pawley, Andrew and Syder, Frances Hodgells. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Richards, Jack and Schmidt, Richard (eds.), *Language and communication*, pp. 191–226. Harlow: Longman.

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44 (1-2): 137–158.

Penadés Martínez, Inmaculada. 2001. ¿Colocaciones o locuciones verbales? *Lingüística Española Actual 23* (1): 57–88.

Pérez Serrano, Mercedes. 2014. Análisis de errores colocacionales en un corpus de aprendientes de ELE. *MarcoELE: Revista de Didáctica de Español como Lengua Extranjera*, 19. Retrieved from: http://marcoele.com/analisis-de-errores-colocacionales-en-un-corpus-de-aprendientes-de-ele/.

Pérez-Paredes, Pascual, Sánchez-Tornel, María and Alcaraz Calero, José M. 2012. Learners' search patterns during corpus-based focus-on-form activities. *International Journal of Corpus Linguistics* 17: 483–516.

Polguère, Alain. 2000. "Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French." In *Proceedings of the Ninth EURALEX International Congress, Volume II*, 517–528. Stuttgart: Universität Stuttgart. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.6350.

Potthast, Martin, Trenkmann, Martin and Stein, Benno. 2010. Netspeak: Assisting writers in choosing words. In Gurrin, Cathal, He, Yulan, Kazai, Gabriella, Kruschwitz, Udo, Little, Suzanne, Roelleke, Thomas, Rüger, Stefan and van Rijsbergen, Keith (eds.), *Advances in Information Retrieval. 32nd European Conference on Information*

Retrieval (ECIR 10), Milton Keynes, UK, March 28-31, 2010, pp. 672-672. Berlin/Heidelberg: Springer.

Real Academia Española (n.d.). *Corpus de referencia del español actual*. Available at: corpus.rae.es.

Read, John. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10 (3): 355–371.

Renau, Irene and Kilgarriff, Adam. 2013. EsTenTen, a vast web corpus of peninsular and American Spanish. *Procedia - Social and Behavioral Sciences* 95: 12–19. Retrieved from: http://www.sciencedirect.com/science/article/pii/S1877042813041372

Revier, Robert Lee. 2009. Evaluating a new test of whole English collocations. In Barfield, Andy and Gyllstad, Henrik (eds.), *Researching collocations in another language*, pp. 125–138. New York: Palgarve Macmillan.

Römer, Ute. 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics* 31: 205–225.

Rundell, Michael. (ed.). 1997. *Longman essential activator*. Harlow, Essex: Longman.

Rundell, Michael. 2010. *Macmillan Collocations Dictionary*. Oxford: Macmillan.

Rychlý, Pavel. 2008. A lexicographer-friendly association score. In Sojka, Petr and Horák Aleš (eds.), *Proceedings in Recent Advances in Slavonic Natural Language Processing*, pp. 6–9. Brno: Masaryk University.

Sanromán Vilas, Begoña. 2003. Semántica, sintaxis y combinatoria léxica de los nombres de emoción en español. (Doctoral dissertation). University of Helsinki, Helsinki. Retrieved from: https://helda.helsinki.fi/handle/10138/19282

Schmitt, Norbert. 2001. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, Norbert, Grandage, Sarah and Adolphs, Svenja. 2004. Are corpus-derived recurrent clusters psycholinguistically valid?” In Schmitt, Norbert (ed.), *Formulaic* *sequences: Acquisition, processing and use*, pp. 127–151. Amsterdam/Philadelphia: John Benjamins.

Serrano-Dolader, David. 2007. ¿Cuál es la utilidad de un 'diccionario combinatorio' de español en el ámbito del español como lengua extranjera? (Echar las REDES… para recoger los frutos). In Luque Toro, Luis (ed.), *Léxico español actual: Actas del I Congreso Internacional de Léxico Español Actual, Venecia-Treviso, 14-15 de marzo de 2005,* pp. 275–283. Venice: Universitá Ca' Foscari Venezia.

Sinclair, John. 1966. Beginning the study of lexis. In Bazell, Charles E., Catford, C., Halliday, Michael A.K. and Robins, R.H. (eds.), *In Memory of J. R. Firth*, pp. 410-430. London: Longman.

———. 1987. Collocation: A progress report. In Steele, Ross and Threadgold, Terry (eds.), *Language topics: Essays in honour of Michael Halliday, Volume 2*, pp. 319–331. Amsterdam: John Benjamins.

———. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

———. 2008. *Collins COBUILD advanced dictionary*. *6th edition*. Boston/Glasgow: Heinle Language Learning/Harper Collins Publishers.

Siyanova, Anna and Schmitt, Norbert. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review/La revue canadienne des langues vivantes* 64 (3): 429–458.

Steele, James and Meyer, Ingrid. 1990. Lexical functions in an explanatory and combinatorial dictionary: Kinds, descriptions, and English examples. In Steele, James (ed.), *Meaning-Text Theory: Linguistics, lexicography and implications*, pp. 41–61. Ottawa: Ottawa University Press.

Suñer Gratacós, Avel·lina. 1999. La aposición y otras relaciones de predicación en el sistema nominal. In Bosque, Ignacio and Demonte, Violeta (eds.), *Gramática descriptiva de la lengua española, Vol. 1.*, pp. 523–564. Madrid: Espasa-Calpe.

Taguchi, Naoko. 2008. Building language blocks in L2 Japanese: Chunk learning and the development of complexity and fluency in spoken production. *Foreign Language Annals* 41 (1): 132–156.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language*. London: Harvard University Press.

Tono, Yukio, Satake, Yoshiho and Miura, Aika. 2014. The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL* 26 (2014): 147–162.

Tutin, Agnès. 2013. Les collocations lexicales: une relation essentiellement binaire définie par la relation prèdicat-argument. *Langages* 189: 47–63.

Tutin, Agnès, Novakova, Iva, Grossmann, Francis and Cavalla, Cristelle. 2006. Esqisse de typologie des noms d'affect à partir de leurs propriétés combinatoires. *Language Française* 150: 32–49.

University of Wisconsin. 1998. *The University of Wisconsin College-Level Placement Test: Spanish (grammar) [Form 96M]*. Madison, WI: University of Wisconsin Press.

Uriel Domínguez, Meritxell. 2014. Las colocaciones en un corpus de aprendices valón y flamenco. (Master's thesis). Universitat de Barcelona/Universitat Pompeu Fabra, Barcelona.

Val Álvaro, José Francisco. 1999. La composición. In Bosque, Ignacio and Demonte, Violeta (eds.), *Gramática descriptiva de la lengua española, Vol. 3.*, pp. 4757–4842. Madrid: Espasa-Calpe.

Vázquez Veiga, Nancy. 2014. "Las marcas de uso en el Diccionario de Colocaciones del Español. *Zeitschrift für romanische Philologie* 130 (3): 698–724.

Verlinde, Serge, Leroyer, Patrick and Binon, Jean. 2009. Search and you will find. From stand-alone lexicographic tools to user driven task and problem-oriented multifunctional leximats. *International Journal of Lexicography* 23 (1): 1–17.

Verlinde, Serge and Peeters, Geert. 2012. Data access revisited: The Interactive Language Toolbox. In Granger, Sylviane and Paquot, Magali (eds.), *Electronic lexicography*, pp. 147–162. Oxford: Oxford University Press.

Verlinde, Serge, Selva, Thierry and Binon, Jean. 2005. Dictionnaires électroniques et environnement d'apprentissage du lexique. *Revue Française de Linguistique Appliquée* X (2): 19–30.

Vincze, Orsolya and Alonso Ramos, Margarita. 2013. Incorporating frequency information in a collocation dictionary: Establishing a methodology. *Procedia - Social and Behavioral Sciences* 95: 241–248. Retrieved from: http://www.sciencedirect.com/science/article/pii/S1877042813041645

Vincze, Veronika. 2005. Funkcióigés szerkezetek vizsgálata lexikai függvények segítségével. Presented at *Nyelvészdoktoranduszok 9. Országos Konferenciája, Szeged, 2005. november 18*.

Vinogradov, Viktor V. 1947. Ob osnovnuikh tipakh frazeologicheskikh edinits v russkom yazuike. In Shakhmatov, Alexei A. (ed.), *1864-1920. sbornik statey i materialov*, 339–364. Moscow: Nauka.

Wanner, Leo. 1996. Introduction. In Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, pp. 1–35. Amsterdam/Philadelphia: John Benjamins.

———. 2006. "¿El corpus como un diccionario de colocaciones?" In Alonso Ramos, Margarita (ed.), *Diccionarios y fraseologia*, pp. 161–173. A Coruña: Servizo de Publicacións Universidade da Coruña.

Wanner, Leo, Alonso Ramos, Margarita, Vincze, Orsolya, Nazar, Rogelio, Ferraro, Gabriela, Mosqueira, Estela and Prieto, Sabela. 2013. Annotation of collocations in a learner corpus for building a learning environment. In Granger, Sylviane, Gilquin, Gaëtanelle and Meunier, Fanny (eds.), *Twenty years of learner corpus research: Looking back, moving ahead*, pp. 493–504. Louvain-la-Neuve: Presses Universitaires de Louvain.

Wanner, Leo, Verlinde, Serge and Alonso Ramos, Margarita. 2013. Writing assistants and automatic lexical error correction: Word combinatorics. In Kosem, Istok, Kallas, Jelena, Gantar, Polona, Krek, Simon, Langemets, Margit and Tuulik, Maria (eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, pp. 472–487. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. Retrieved from: http://eki.ee/elex2013/proceedings/eLex2013_33_Wanner+Verlinde+Alonso-Ramos.pdf

Wible, David and Tsao, Nai-Lung. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions.” In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, 25–31. Los Angeles, CA: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/W/W10/W10-0804

Williams, Geoffrey. 1998. Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3 (1): 151–171.

Willis, Dave. 1990. *The Lexical SyIlabus*. London/Glasgow: Collins E.L.T.

———. 1999. Syllabus design and the pedagogic corpus. In Aitchison, Jean et al. (eds.), *Vocabulary learning in a foreign language*, pp. 115–148. Fontenay/St.-Cloud: ENS Editions.

Witten, Ian H., Wu, Shaoqun, Li, Liang and Whisler, Jennifer L. 2013. *The book of FLAX: A new approach to computer-assisted language learning. Second Edition.* Hamilton, New Zealand.

Wood, David. 2009. Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics* 12 (1): 39–58.

Woolard, George. 2000. Collocation - Encouraging learner independence. In Lewis, Michael (ed.), *Teaching collocation: Further developments in the Lexical Approach*, pp. 28–46. Hove: Language Teaching Publications.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21 (4): 463–489.

———. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wu, Jian-Cheng C, Chang, Yu-Chia., Mitamura, Teruko and Chang, Jason S. 2010. Automatic collocation suggestion in academic writing. *Proceedings of the ACL 2010 Conference Short Papers*, pp. 115–119. Uppsala: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/P10-2021

Wu, Shaoqun. 2010. Supporting collocation learning. (Doctoral dissertation). University of Waikato, Hamilton. Retrieved from: http://researchcommons.waikato.ac.nz/handle/10289/4885

Wu, Shaoqun, Franken, Margaret and Witten, Ian H. 2010. Supporting collocation learning with a digital library. *Computer Assisted Language Learning* 23 (1): 87–110.

Wu, Shaoqun, Witten, Ian H. and Franken, Margaret. 2010. Utilizing lexical data from a web-derived corpus to expand productive collocation knowledge. *ReCALL* 22 (01): 83–102.

Yoon, Hyunsook. 2008. More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology* 12 (2): 31–48.

Yoon, Hyunsook and Hirvela, Alan. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing* 13 (4): 257–283.

Yorio, Carlos A. 1989. Idiomaticity as an indicator of second language proficiency.” In Hyltenstam, Kenneth and Obler, Loraine K. (eds.), *Bilingualism across the lifespan:*

Aspects of acquisition, maturity and loss, pp. 55–72. Cambridge: Cambridge University Press.

Yuldashev, Aziz, Fernández, Julieta and Thorne, Steven L. 2013. Second language learners' contiguous and discontiguous multi-word unit use over time. *The Modern Language Journal* 97 (51): 31–45.

Zaferanieh, Elaheh and Behrooznia, Saeedeh. 2011. On the impacts of four collocation instructional methods: Web-based concordancing vs. traditional method, explicit vs . implicit instruction. *Studies in Literature and Language* 3 (3): 120–126.

Zholkovsky, Aleksandr K. and Mel'čuk, Igor. 1970. Sur la synthèse sémantique. *T. A. Informations* 2: 1–85.

———. 1984. *Explanatory combinatorial dictionary of modern Russian*. Vienna: Wiened Slawistischer Almanach.

Appendix A. DiCE usability test: Original version

El Diccionario de Colocaciones del Español (DiCE) está accesible en la web desde hace tiempo. El objetivo de este cuestionario es ver hasta qué punto resulta transparente la interfaz de búsquedas del diccionario a los usuarios.

1. Consulta del DiCE

Accede a la página web del DiCE (www.dicesp.com), toma tu tiempo para conocer el diccionario y la interfaz, lee la descripción del diccionario y de las búsquedas si es necesario. Cuando crees que estás preparado/a, intenta contestar las siguientes preguntas, haciendo uso de las opciones de búsqueda. Después de cada búsqueda indica en la escala 1-5 el nivel de dificultad en/para encontrar la información en el diccionario (1=muy fácil; 5=imposible).

Mi IP: ________________________

Conexión: desde _______________ hasta________________

Antes de empezar verifica tu dirección IP en www.whatsmyip.org y anótalo junto a la hora exacta de tu conexión en el siguiente recuadro:

IMPORTANTE: Después de finalizar cada búsqueda, tienes que volver a la **pantalla de inicio** (***Bienvenida***) del diccionario. Esto nos sirve para poder medir el tiempo que pasas haciendo cada búsqueda.

1. ¿Cuáles son los verbos que se pueden combinar con la unidad léxica ***cariño 2***? Escríbelos en el siguiente recuadro.

__________ cariño

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

2. ¿Cuál es el significado de la combinación ***reanudar la amistad***? Escribe en el recuadro la glosa que figura en el DiCE.

____

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

3. Busca el/los adjetivo(s) adecuado(s) para expresar el sentido ‘amor que se tienen uno al otro’. Escribe el/los adjetivo(s) en el recuadro.

Amor __________

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

4. ¿Significa lo mismo ***captar la atención*** que ***centrar la atención***? Escribe en el recuadro el significado de estas colocaciones según el DiCE.

____

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

5. Busca todas las colocaciones del diccionario que tengan como colocativo el verbo ***cumplir*.** Escribe en el recuadro los nombres que se combinan con ese verbo

cumplir __________

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

6. ¿Qué adjetivos se pueden utilizar para hablar de ***celos*** 'demasiado intensos'? Escríbelos en el siguiente recuadro.

celos __________

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

7. ¿Cuáles son las colocaciones de la palabra ***remordimiento*** que están codificadas por la función léxica Sing? Escribe en el recuadro los colocativos.

__________ remordimiento

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *BIENVENIDA******

8. Busca todas las colocaciones en el DiCE que tengan como colocativo el adjetivo ***negro,*** bien en la forma del masculino o bien en la del femenino. Escribe las bases en el recuadro.

__________ negro/a

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *<u>BIENVENIDA</u>******

9. ¿Cuál es la colocación con la palabra ***rencor*** que está codificada por la función léxica Magn+A2Manif? Escribe el colocativo en el recuadro.

___________ rencor

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *<u>BIENVENIDA</u>******

10. Lee la siguiente frase:

Los agresores del pequeño suelen ser los familiares o <u>amistades</u> más cercanas.

Intenta identificar con qué sentido, es decir, con qué unidad léxica del nombre *amistad* coincide en el DiCE. Escribe todos los adjetivos que pueden coocurrir con *amistad* en ese sentido. Ten cuidado, debes indicar únicamente los adjetivos que figuran en la entrada del DiCE correspondiente a esa unidad léxica.

____

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *<u>BIENVENIDA</u>******

11. ¿Se encuentra en el DiCE la combinación ***honda felicidad***?

Sí No

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *<u>BIENVENIDA</u>******

12. Completa la siguiente frase con los adjetivos descritos por la glosa *grande*. Indica solo aquellos adjetivos que pueden aparecer según el DiCE junto a la unidad léxica de ***conmoción*** cuyo significado corresponde con la frase.

Tuvo que ser trasladado de urgencia a un centro médico por sufrir una ____________ (grande) conmoción.

Dificultad: 1 2 3 4 5

*****VUELVE A LA PANTALLA *<u>BIENVENIDA</u>******

13. ¿Qué preposición debemos usar con la colocación ***sembrar miedo*** ______ ***alguien***? Completa la siguiente frase:

> En alta mar, el famoso monstruo sembraba miedo ________ los marineros.

Dificultad: 1 2 3 4 5

VUELVE A LA PANTALLA BIENVENIDA

2. Tu opinión sobre el DiCE

a) ¿Sueles utilizar diccionarios electrónicos? ¿Cuáles?

b) ¿Has utilizado el DiCE anteriormente?

c) ¿Utilizas el DiCE con frecuencia?

d) ¿Crees que el DiCE es una herramienta útil? ¿Para qué tipo de usuarios crees que resulta más útil? ¿Tú para qué lo utilizas o utilizarías?

e) ¿Volverás a utilizar el DiCE?

f) ¿Recomendarías a otras personas que lo utilicen?

g) ¿Cómo crees que se podría mejorar?

h) ¿Hay alguna consulta que te gustaría poder hacer y actualmente no es posible?

Diccionario de Colocaciones del Español (DiCE) has been available on the web for some time. The aim of this questionnaire is to find out to what extent users find the query interface to be transparent.

Appendix B. DiCE usability test: English version

The Diccionario de Colocaciones del Español (DiCE) has been available on the web for a long time. The purpose of this questionnaire is to evaluate to what extent the dictionary interface and query options are clear to the users.

1. Querying DiCE

Enter the DiCE web site (www.dicesp.com), take your time to get to know the dictionary and its interface, read the information provided on the dictionary and the search options if necessary. Once you think you are ready, try to answer the following questions using the different search options. After each query, indicate on a 1-5 scale how difficult it was to find the information (1=ver easy; 5=imposible).

Before you start, check your IP address at www.whatsmyip.org and write it together with the hour and date of entry in the following box:

Mi IP: ___________________________

Conexión: desde hasta

IMPORTANT: After finishing each query, you have to return to the **main page *(Bienvenida)*** of the dictionary. This is important since it helps us measure the time you spend on each query.

1. What verbs can be combined with the lexical unit ***cariño 2*** ‘affection’? Write them in the following box.

__________ cariño

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

2. What is the meaning of the combination ***reanudar la amistad*** ‘renew a friendship’? Write the gloss shown in DiCE in the following box.

____

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

3. Search for the suitable adjective(s) to express the meaning ‘love felt for one another’. Write it/them in the box

Amor __________

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

4. Do ***captar la atención*** ‘catch somebody’s attention’ and ***centrar la atención*** ‘focus somebody’s attention’ have the same meaning? Write in the box the meaning of these collocations according to DiCE.

____

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

5. Search for all the collocations in the dictionary containing the collocate verb ***cumplir*** ‘fulfill’. Write the nouns which can be combined with this verb in the box.

cumplir __________

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

6. What adjectives can be used to speak about “extreme” ***celos*** ‘jealousy’? Write them in the box.

celos __________

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

7. What are the collocations of the word ***remordimiento*** ‘remorse’ which are encoded by the lexical function Sing? Write the collocates in the box.

__________ remordimiento

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

8. Find all possible collocations in DiCE containing the adjective collocate ***negro*** ‘black’ both in the masculine or the feminine form. Write them in the box.

__________ negro/a

Difficulty: 1 2 3 4 5

*****RETURN TO PAGE *BIENVENIDA******

9. What collocation of the word ***rencor*** 'grudge' is encoded by the lexical function Magn+A2Manif? Write the collocate in the box.

__________ rencor

Difficulty: 1 2 3 4 5

RETURN TO PAGE *BIENVENIDA*

10. Read the following sentence:

Los agresores del pequeño suelen ser los familiares o amistades más cercanas.
'Assailants of the child are often family members or close acquaintances.'

Try to identify to which sense, in other words, to which lexical unit of the noun *amistad* 'friendship' it corresponds in DiCE. Write all the adjectives which can co-occur with this lexical unit of *amistad* in the box. Be careful to only list the adjectives appearing in DiCE in the entry of the given lexical unit.

____

Difficulty: 1 2 3 4 5

RETURN TO PAGE *BIENVENIDA**

11. Is the combination ***honda felicidad*** 'profound happiness' included in DiCE?

Yes No

Difficulty: 1 2 3 4 5

RETURN TO PAGE *BIENVENIDA**

12. Complete the following sentence with adjectives described by the gloss *grande* 'great'. Include only the adjectives which according to DiCE can co-occur with the lexical unit of ***conmoción*** 'concussion' corresponding to the sentence.

Tuvo que ser trasladado de urgencia a un centro médico por sufrir una ____________ (grande) conmoción.

'It had to be an emergency transfer to a medical center due to

Difficulty: 1 2 3 4 5

RETURN TO PAGE *BIENVENIDA**

13. What preposition should be used with the combination ***sembrar miedo ___ alguien*** 'to instil fear in somebody'? Complete the following sentence:

En alta mar, el famoso monstruo sembraba miedo ________ los marineros. 'On the open sea, the famous monster

Difficulty: 1 2 3 4 5

RETURN TO PAGE *BIENVENIDA**

2. Your opinon on DiCE

a. Do you use electronic dictionaries? Which ones?

b. Had you used DiCE before?

c. Do you use DiCE frequently?

d. Do you think that DiCE is a useful resource? For what kind of users do you think it could be especially useful? What would you use it for?

e. Will you use DiCE again?

f. Would you recommend DiCE to others?

g. How could DiCE be improved?

h. Is there any query you would like to be able to do and is not available currently?

Appendix C. Collocation error correction with concordances: Combined version of the two types of original questionnaires including both n-gram and full sentence concordance feedback for test items concerned

Datos personales:

El cuestionario es anónimo, pero para poder sacarles más provecho a los resultados, necesitamos que nos proporciones algunos datos personales.

sexo: masculino femenino

edad: …………….

país: …………….

lengua(s) materna(s): …………….…………….…………….…………….…………….

lengua(s) materna(s) de tu madre: …………….…………….…………….……………..

lengua(s) materna(s) de tu padre: …………….…………….…………….……………...

¿Qué lengua(s) hablas en casa? …………….…………….…………….…………….

¿Cuánto tiempo llevas estudiando español? …………….…………….……………........

¿Cuál es tu nivel? …………….…………….…………….…………….........................

¿Has vivido en un país hispanohablante? …………….……………...............................

¿Dónde y cuánto tiempo? …………….…………….…………….…………….……….

¿Qué otros idiomas hablas? ¿Qué nivel tienes?

idioma: ……………. nivel: …………….

……………. …………….

……………. …………….

Ejercicio 1
Las siguientes frases fueron escritas por estudiantes de español. Los segmentos resaltados son erróneos. ¿Puedes corregirlos? **NO** puedes utilizar ningún tipo de ayuda (diccionario, libro de gramática). Escribe tu corrección debajo de la frase.

1. Quiero precisar que, aunque el libro es cuasi-histórico, contiene una sarta de **misenterpretaciones**, errores y saltos de fantasía.

2. Ahora tengo nuevas cosas para ocuparme que **me ponen muy apasionada**, pero tampoco me dan una idea más clara de cómo será mi futuro.

3. Basta con estar en una parada **esperando un metro** para ver a los fumadores disfrutando de un cigarrillo mientras los demás intentan huir del humo.

4. Cuando alguien dice que los gays no deben **tener los derechos para casarse**, me enfada mucho porque la razón siempre es la religión o la ignorancia de la persona.

5. Cuando la madre se enteró, estaba enfadadísima, y al buscar a su marido lo encontró con la madre de Agustina, y entonces **encendió el fuego que quemó la casa** y los mató.

6. Me gustaría **montar una bicicleta** en el bosque tropical.

7. Durante el verano trabajo en un campamento para niños de distintos países. El año pasado, después del campamento y a punto de **volver loco** después de tanto tiempo rodeado de niños, fui a Barcelona con mi novia para descansar.

8. Al día siguiente, nos despedimos, y **gracias,** y caminamos hacia el puerto.

9. En cuanto al **futuro lejos**, también tengo muchas ideas. Quiero trabajar en una universidad o enseñar inglés para extranjeros o enseñar cómo enseñar inglés.

10. La familia decide veranear en Malibú y quieren que la empleada vaya con ellos. Ella no acepta ir por su hija y **usa excusas** como, "no quiero trabajar tan lejos de mi hija."

11. La hija está tratando de **capturar la atención** de su madre, pero es muy difícil porque la madre siempre trabaja.

12. **La película se trata de una mujer soltera**, su hija y sus amigas (casi no hay hombres en la película).

13. **Las temperaturas frescan un poco** y, al mismo tiempo, las hojas de los árboles cambian también.

14. **Los derechos mujeriles** empezaban a mejorar en casi todos los regiones del mundo, aún hasta tener chancilleres mujeres en varios países hoy en día.

15. En la cena **dimos bienvenidas a los nuevos estudiantes**, y dijimos adiós a los que se iban.

16. Mi futuro **no tiene limitades**. Ahora solamente tengo veinticinco años.

17. No creo que mi opinión vaya a **cambiar la mente de quienes** que han dicho que los gays no deben tener derechos.

18. No hay que olvidar que hay problemas en el barrio. Hay personas que venden drogas, y a veces hay **crímenes violentas**.

19. Si los africanos se sintieran más como españoles verdaderos, es posible que **les dé la gana de** aprender más el idioma.

20. Siempre hay gente sentada en sillas en la explanada, **haciendo de cotilleo** o leyendo el periódico.

Ejercicio 2
Ahora corrige los segmentos resaltados de las frases con la ayuda de los ejemplos que aparecen junto a cada una. **NO** puedes utilizar ningún otro tipo de ayuda (diccionario, libro de gramática). Escribe tu corrección debajo de la frase.

1. Quiero precisar que, aunque el libro es cuasi-histórico, contiene una sarta de **misenterpretaciones**, errores y saltos de fantasía.

El descubrimiento de la circulación sanguínea provocó algunas interpretaciones *erróneas* curiosas.
Es vital ser específico en esto, si queremos evitar interpretaciones *erróneas*.
Ellos reconocían su importancia para una interpretación *correcta* de la realidad.
Es prácticamente imposible cometer errores de medición que se deriven de una interpretación *correcta* de las indicaciones.
Esta creencia está basada en interpretaciones *incorrectas* apoyadas en evidencias dudosas.
Tu respuesta ha sido sin fundamento, mal basada en tus impulsos emocionales y una interpretación *incorrecta* de mi comentario.
Como también se silencia la "discriminación de los discapacitados" que se produce a raíz de una mala interpretación del diagnóstico prenatal.
Lo digo para que no haya lugar a malas interpretaciones.

2. Ahora tengo nuevas cosas para ocuparme que **me ponen muy apasionada**, pero tampoco me dan una idea más clara de cómo será mi futuro.

Cualquier asignatura bien enseñada puede *despertar* pasiones.
Podemos asegurar que el grupo ha *despertado* pasiones entre el público adolescente.
Los equipos de fútbol *levantan* pasiones y tienen vuelos muy a menudo por los diversos partidos que disputan cada semana en diferentes territorios.
La prueba de que *levanta* pasiones es la cantidad de comentarios que se pueden leer sobre él y su programa solo en este foro.
Siente pasión por sus orígenes y es el mejor embajador segoviano del norte peninsular.
Isidoro y los voluntarios del programa *sienten* verdadera pasión por su comarca.
Me apasiona el reto de mejorar la calidad de vida y considero que cualquier persona puede conseguirlo si se lo propone.
Espero que tengas mucho trabajo y muchos éxitos en esto que tanto te apasiona .

3. Basta con estar en una parada **esperando un metro** para ver a los fumadores disfrutando de un cigarrillo mientras los demás intentan huir del humo.

Colocar máquinas donde los visitantes pueden recoger libros para leer mientras *esperan* el metro y soportar las largas colas.
Mientras *esperan* el metro , pueden sentirse en la república independiente de su casa.
Ya no necesitaremos *coger* el último metro o esperar al primero.
Quedamos a las seis de la tarde para *coger* el metro a la playa de la que tan solo nos separa una estación.
Tuve que subir varias veces al metro.
Si vais a Tokyo no dudéis en subir al metro.

4. Cuando alguien dice que los gays no deben **tener los derechos para casarse**, me enfada mucho porque la razón siempre es la religión o la ignorancia de la persona.

Full sentence concordances:
El cónyuge y los hijos a cargo *tienen* derecho a ejercer una actividad económica.
Los abonados *tendrán* derecho al uso libre de gimnasio y piscina.
El comité organizador se *reserva* el derecho de decidir cualquier aspecto de la competición que no esté recogido en estas bases.
Barclays se *reserva* el derecho a modificar o suprimir las condiciones de esta oferta o a sustituir las empresas oferentes en cualquier momento y sin previo aviso.
Fecha a partir de la cual las nuevas acciones *dan* derecho a participar en las ganancias sociales.
También *darán* derecho a 1 crédito de libre configuración para aquellos interesados.

N-gram concordance format:
ni tiene el derecho de
tiene derecho todo el
tienen derecho todos los
tener los mismos derechos de
tendrá además derecho para que
tendrán derecho todos los
tiene derecho y que no
tenían el derecho de

tenemos el derecho y la
y no tiene derecho de
tienen este derecho y el

5. Cuando la madre se enteró, estaba enfadadísima, y al buscar a su marido lo encontró con la madre de Agustina, y entonces **encendió el fuego que quemó la casa** y los mató.

Full sentence concordances:

Posteriormente se informó además de que algunos manifestantes habían *prendido* fuego a un edificio en el centro de la capital.
Resulta que al parecer intentó *prender* fuego a su mujer rociándola con gasolina.
Los trabajadores de la empresa vecina nos dieron un bidón con palos para que *hiciéramos* fuego.
Nunca *hagas* fuego en el campo.
Encender fuego solamente en lugares autorizados y acondicionados para ello.
La identidad del encargado de *encender* el fuego olímpico es normalmente mantenida en secreto hasta último momento.

N-gram concordance format:

para prender el fuego de
enciende el fuego de la
para encender el fuego de
de encender el fuego y
encendiendo el fuego de la
prender fuego a todos los
y prendieron fuego a la
y luego prendieron fuego a
y prendió fuego a las
y prender fuego a los

6. Me gustaría **montar una bicicleta** en el bosque tropical.

Full sentence concordances:

Utilizo la bicicleta para todo lo que no me implica llevar carga o desplazarme a una distancia demasiado larga.
Personalmente *utilizo* la bicicleta a menudo para hacer la compra y salir por la noche.
Yo vivo solo y puedo *coger* la bici para hacer todo tipo de cosas.
Con eso ya consigues que baje la intensidad del tráfico y más gente se atreverá a *coger* la bici.
Decirte también que en cardio estoy *haciendo* bicicleta elíptica e intento meter intervalos de alta intensidad.
Pero ella está dispuesta a escalar, a *hacer* bicicleta de montaña y a practicar el ala delta sin miedo a destacar.
Montar en bicicleta es un ejercicio relajante además de divertido.
Yo me considero una novel en esto de montar en bici.

N-gram concordance format:

montarse en el tren de
y montó en el coche
montaron en el helicóptero y
montarse en el carro de
montó en su bicicleta y
montaron en el coche y
montaron en el carro y
montó en el carro y
monté en el coche y
montó en su coche y
montó en la bicicleta y

7. Durante el verano trabajo en un campamento para niños de distintos países. El año pasado, después del campamento y a punto de **volver loco** después tanto tiempo rodeado de niños, fui a Barcelona con mi novia para descansar.

Full sentence concordances:

Tiene el respeto de sus compañeros y no es fácil *volverlo* loco.
Morenas y rubias con impresionantes medidas *vuelven* locos a los hombres de medio mundo.
Así que me alegro de que empiece el curso antes de que empiece a *volverme* loca.
Así que si quieres venir a visitar la isla no te *vuelvas* loco intentando reservar en vuelos directos!
El tiempo libre era lo más deseado por los chavales que *estaban* locos por meterse en el mar.
Y es que pensaba que mis compañeros de viaje *estaban* locos por querer ponerse a esperar más de ocho horas a que abrieran las puertas del recinto.

N-gram concordance format:

vuelve locos a los
volvía locas a las
para volver loco a cualquier
vuelve locas a las mujeres

vuelve loca a la gente
volver loco a cualquier hombre
se volvió loco después de
se volvió loco de tanto
me volviese loco de aquel
se vuelven locos por el
se volvió loco por la
se vuelven locos por

8. Al día siguiente, nos despedimos, y **gracias,** y caminamos hacia el puerto.

Además ***doy*** las **gracias** a esos internautas que pierden el tiempo en subirlos.
Desde aquí le ***damos*** las **gracias** por su buena disposición.
De antemano te ***digo*** **gracias**.
A todos os queremos ***decir*** **gracias** de corazón.

9. En cuanto al **futuro lejos**, también tengo muchas ideas. Quiero trabajar en una universidad o enseñar inglés para extranjeros o enseñar cómo enseñar inglés.

La aplicación en un **futuro *próximo*** la quiero liberar aunque hasta eso tengo que hacerla un poco más amena al usuario.
La idea podría adaptarse a un juguete para niños aunque la compañía no tiene previsto comercializarlo en un **futuro *próximo***.
Preguntar y hablar sobre planes en un **futuro *inmediato***.
No es momento de comprometer el **futuro *inmediato*** con "ocurrencias electorales".
Muchas distribuidoras solamente están interesadas en el **futuro *cercano***.
Otros países están interesados en este programa y podrían también aplicarla como norma en un **futuro *cercano***.
Pero tal y como funciona el mundo lo mal que está dudo que esto cambie para mejor y esto será en un **futuro *lejano*** y no cercano.
Protege el planeta de los alienígenas y otros enemigos que llegan de un **futuro *lejano***.

10. La familia decide veranear en Malibú y quieren que la empleada vaya con ellos. Ella no acepta ir por su hija y **usa excusas** como, "no quiero trabajar tan lejos de mi hija."

Esta vez ningún vecino va a ***tener*** la **excusa** de decir que no se ha enterado.
Así que los mayores fans ya ***tienen*** una **excusa** para recorrer ese fantástico país.
Es cierto que las reglas sobre los resultados de la votación eran muy precarias y no cabe ***buscar*** **excusas** ni culpables.
Pero a veces nos ***buscamos*** **excusas** para no ir al gimnasio.
Entonces ***puse*** la **excusa** de que tenía pocas amigas y estaba en edad de salir con chicas.
Ahora ya no podemos ***poner*** la **excusa** de que no sabemos dónde está la biblioteca más próxima.
Que no ***dé*** más **excusas** de mal pagador.
Los juguetes te ***darán*** la **excusa** perfecta para pasar un largo rato con tu gato darle un mimo merecido a la mascota de la casa.

11. La hija está tratando de **capturar la atención** de su madre, pero es muy difícil porque la madre siempre trabaja.

Full sentence concordances:

La ingesta de líquidos es a lo que habitualmente ***prestamos*** menos **atención**.
Los niños sienten celos y los expresan a través de necesidades a las que los padres deben ***prestar*** **atención**.
Esta siempre ha sido la enfermedad que más me ha asombrado y que me ha ***llamado*** la **atención** de una forma especial.
Aquí os vamos a describir algunas opciones que nos han ***llamado*** la **atención** por sus cualidades.
Connelly ***atrae*** la **atención** y no deja que decaiga en ningún momento.
El otro día estaba comprándome libros en la librería y este me ***atrajo*** la **atención**.
Desarrollar una cierta habilidad para el coqueteo te puede ayudar a aprender a ***captar*** la **atención** e interés de las personas que te agradan.
Los niños suelen portarse mal para ***captar*** la **atención** de los padres.

N-gram concordance format:

capten la atención de
captar la atención y la
para captar la atención de
concentrado la atención de los
concentra la atención de
de concentrar la atención de
capten la atención de los
concitan la atención de los

concitaban la atención de los
captan la atención de los

12. **La película se trata de una mujer soltera**, su hija y sus amigas (casi no hay hombres en la película).

Full sentence concordances:

La **película** ***trata*** de un psicópata disfrazado de samurai que droga y despedaza lentamente a una mujer.
La **película** ***trata*** de una supuesta herencia multimillonaria que le llega a una familia.
La **película** ***muestra*** con crudeza los efectos de la enfermedad en las personas.
La **película** ***muestra*** un tema tan apasionante como sensible a ciertas susceptibilidades.
La **película** ***narra*** la historia de 3 jóvenes que se pierden en un bosque mientras filman un documental sobre una leyenda local.
La **película** ***narra*** tres historias relacionadas con la muerte que finalmente se cruzan.

N-gram concordance format:

películas que tratan de la
las películas que traten de
una película que trata de
las películas que tratan de
una película que trataba de

13. **Las temperaturas frescan un poco** y, al mismo tiempo, las hojas de los árboles cambian también.

Sin embargo hay que señalar que las **temperaturas** ***descienden*** considerablemente en la altura.
Por el contrario si la **temperatura** ***desciende*** el aire se contrae empujando al otro índice hasta la posición de mínima.
Las **temperaturas** ***bajarán*** para el fin de semana.
Las **temperaturas** ***bajarán*** entre dos y tres grados respecto a las del jueves.
El domingo, las **temperaturas** comenzarán a ***refrescarse*** y sólo quedarán en ese nivel de alerta siete provincias.
La **temperatura** se ha ***refrescado*** y ha de ser seguramente porque los árboles de sombra han espesado.

14. **Los derechos mujeriles** empezaban a mejorar en casi todos los regiones del mundo, aún hasta tener chancilleres mujeres en varios países hoy en día.

Democracia y los **derechos** ***humanos***.
Lo cierto es que lo interesante sería tener **derechos** ***humanos*** para los humanos.
El trabajo es un **derecho** de la ***mujer*** actual que le permite ser independiente económicamente y realizarse como persona.
Si la ley vigente es éticamente inadmisible, «por lo menos no trata al aborto como un **derecho** de la ***mujer***».

15. En la cena, donde **dimos bienvenidas a los nuevos estudiantes**, y dijimos adiós a los que se iban.

Full sentence concordances:

La ciudad ***dará*** la **bienvenida** a los estudiantes que vivirán dos días de intercambio en los que el sushi y el manga compartirán mesa con las tapas y el flamenco.
Nos ***dio*** la **bienvenida**.
El personal y los dueños son muy amables y siempre ***ofrecen*** una cálida **bienvenida** a sus clientes.
El equipo me ha ***ofrecido*** una gran **bienvenida** y todo el mundo me ha ayudado, ya que sabían que era la primera vez que pilotaba el coche.

N-gram concordance format:

dar bienvenida a
da la bienvenida a sus
dan la bienvenida a
dan la bienvenida a este
dar bienvenida en
dieron la bienvenida en el
darle la bienvenida en esta
da la bienvenida en el
nos dieron la bienvenida y
le da la bienvenida como
les daba la bienvenida y
le doy la bienvenida y

16. Mi futuro **no tiene limitades**. Ahora solamente tengo veinticinco anos.

Pero la obsesión no ***tiene*** **límites**.
La codicia de los políticos para arrancar dinero de nuestro esfuerzo no ***tiene*** **límites**.
El análisis de todo ello nos acercará más a la realidad pero ***tiene*** serias **limitaciones** relacionadas con la recogida de los datos y su naturaleza.

Sin embargo, es una técnica que puede resultar arriesgada ya que ***tiene*** grandes **limitaciones**.

17. No creo que mi opinión vaya a **cambiar la mente de quienes** han dicho que los gays no deben tener derechos.

Han logrado cambiar la *mentalidad* de sus compañeros.
Se ha hecho una inversión tremenda en tecnología antes de cambiar la *mentalidad* del docente".
Hay que cambiar la *idea* de esa necesidad de tener un hijo.
Sin pretender cambiar las *ideas* de nadie, siempre sumando y nunca restando.
La derrota contra el Cajasol no debe cambiar la *opinión* de nadie sobre el equipo.
Pero un buen trabajo no va a cambiar la *opinión* que tengo de los otros tres.
Es conveniente cambiar de entrenador a estas alturas y así cambiaría la *actitud* perdedora del equipo.
Vale que esto es mundial pero como este tio no cambie la *actitud* no se yo hasta donde llegaremos.

18. No hay que olvidar que hay problemas en el barrio. Hay personas que venden drogas, y delincuencia, y a veces **crímenes violentas**.

Las iniciativas de nuestros jueces persiguiendo crímenes *atroces* allí donde se han cometido prestigian a nuestro sistema judicial y dignifican a nuestro país.
Queremos saber quiénes son los responsables de estos crímenes *atroces*, queremos saber quién debe ser condenado.
Entre los años 1994 y 2001 el crimen *violento* ha aumentado en un 33% y los asesinatos a agricultores blancos está creciendo constantemente desde 1991.
Todo lo que tienes que hacer es leer los detalles de un crimen *violento* y mirar las fotografías.
Pena que tú no tengas la decencia de incluir 40 años de dictadura fascista en esos "crímenes especialmente *graves*".
Los crímenes más *graves* de trascendencia para la comunidad internacional en su conjunto.

19. Si los africanos se sentirían más como españoles verdaderos, es posible que **les dé la gana de** aprender más el idioma.

No *tenía* ganas de enfrentarme de nuevo a la carretera pero no me quedaba mas remedio.
Tenía ganas de decirle lo que sentía por ella pero no quería estropearlo.
La verdad es que *dan* ganas de entrenar y probar una de estas competiciones.
Había días en que me *daban* ganas de quedarme sentado a mi mesa y dejarles hacer lo que quisieran.
Tendrá la impresión de que está siendo aplastada por el peso de sus responsabilidades y le podrían *entrar* ganas de abandonarlo todo.
Me *entraron* ganas de llorar, pero me aguanté.

20. Siempre hay gente sentado en sillas en la explanada, **haciendo de cotilleo** o leyendo el periódico.

Y ahora les *cuento* el cotilleo de la semana.
Ahora que ha pasado todo os *cuento* un cotilleo que me han contado y que proviene de alguien que ha estado trabajando en Sudáfrica cubriendo el Mundial.
No lo digo para cotillear, lo digo para ver hasta qué punto le limita su estado.
Para mi el tuenti es mejor que el facebook y el tuenti no se utiliza para cotillear.

Appendix D. Collocation error correction with concordances: English translation of instructions

Personal information:

This is an anonymous questionnaire, however, in order to be able to better interpret the results, we need you to provide some personal information.

sex: male female

age: …………….

country: …………….

native language(s): …………….…………….…………….…………….…………….

mother's native language(s): …………….…………….…………….…………….…..

father's native language(s): …………….…………….…………….…………….……

What language(s) do you use at home? …………….…………….…………….……...

How long have you been learning Spanish? …………….…………….…………….…

What is your proficiency level? …………….…………….…………….……………...

Have you lived in a Spanish-speaking country? …………….…………….…………...

Where and for how long? …………….…………….…………….…………….………

What other foreign languages do you speak? What is your proficiency level?

language: ……………. level: …………….

……………. …………….

……………. …………….

Instructions:

Exercise 1
The following sentences were written by learners of Spanish. The underlined segments are incorrect. Can you correct them? You are **NOT ALLOWED** to use any external aid (dictionary, grammar reference book). Write the correct version underneath the sentence.

Exercise 2

Now you have to correct the underlined segments with the help of the examples provided together with each sentence. You are **NOT ALLOWED** to use any other (dictionary, grammar reference book). Write the correct version underneath the sentence.

Appendix E. Collocation errors included in the test on collocation error correction with concordances

No.	Erroneous combination	Expected correction	Error type
1	*misinterpretaciones* 'misinterpretations'	*malas interpretaciones* 'wrong interpretations'	lexical > synthesis
2	*me ponen muy apasionada* lit. 'they make me passionate'	*me apasionan* 'they fascinate me'	lexical > analysis
3	*esperando un metro* lit. 'waiting for a metro'	*esperando el metro* 'waiting for the metro'	gram. > determination
4	*tener los derechos para casarse* lit. 'have the rights to'	*tener (el) derecho de* 'have (the) right to'	gram. > number gram. > preposition governed by the base (gram. > determination)
5	*encendió el fuego que quemó la casa* lit. 'she lit the fire that burnt the house'	*prendió fuego a la casa* 'she set fire to the house'	lexical > erroneous collocate
6	*montar una bicicleta* 'ride a bike'	*montar en (una) bicicleta* lit. 'ride on a bike'	gram. > preposition governed by the collocate (gram. > determination)
7	volver loco 'to go mad'	*volverme loco* 'to go mad-1sg'	gram. > pronombre
8	*gracias* 'thank you'	*dimos las gracias* 'we said thank you'	lexical > synthesis
9	*future lejos* lit. 'far future'	*futuro lejano* 'distant future'	lexical > incorrect collocate
10	*usa excusas* lit. 'she uses excuses'	*pone excusas* 'she makes up excuses'	lexical > incorrect collocate
11	*capturar la atención* lit. 'to capture sb's attention'	*captar la atención* 'to catch sb's attention'	lexical > incorrect collocate
12	*la película se trata de* 'the movie is about'	*la película trata de* 'the movie is about'	gram. > pronoun
13	*las temperaturas frescan* lit. 'the temperatures turn cooler'	*las temperaturas se refrescan* lit. 'the temperatures turn cooler'	lexical > incorrect collocate
14	*derechos mujeriles* lit. 'womanly rights'	*derechos de las mujeres* 'women's rights'	lexical > incorrect collocate
15	*dimos bienvenidas* lit. 'we gave welcomes'	*dimos la bienvenida* lit. 'we gave the welcome' 'we welcomed'	gram. > number gram. > determination
16	*no tiene limitades* 'it has no limits'	*no tiene límites* 'it has no limits'	lexical > incorrect base
17	*cambiar la mente* lit. 'to change sb's mind'	*cambiar la idea/opinión* 'to change sb's idea/opinion'	lexical > incorrect base
18	*crímenes violentas* 'violent crimes'	*crímenes violentos* 'violent crimes'	gram. > gender
19	*les dé la gana de* 'they feel like doing sg'	*tengan ganas de* 'they wish to do sg'/	lexical > correct collocation with incorrect meaning
20	*haciendo de cotilleo* lit. 'making gossip'	*cotilleando* 'gossiping'	lexical > analysis

Appendix F. Resumen de la tesis

Motivación y objeto de estudio

A lo largo de las últimas décadas, ha aumentado el interés con respecto al vocabulario dentro del ámbito de la enseñanza de segundas lenguas (L2). Además, se ha subrayado que el conocimiento léxico no solo se compone de palabras individuales, sino también incluye una cantidad considerable de expresiones multilexémicas. Este punto de vista se ve reflejado en el trabajo de Pawley y Syder (1983, 193) quienes afirman que dentro de las oraciones gramaticales que se pueden formular en una lengua, solo un subconjunto de oraciones serían consideradas por hablantes nativos como expresiones normales, frente al resto que serían calificadas como raras, no naturales o extranjerismos, por no tener en cuenta que las palabras no se combinan libremente. En consecuencia, se defiende que un aprendiz no puede llegar a un dominio de la lengua parecido al de un hablante nativo sin tener un conocimiento suficiente de las unidades multipalabra, por lo que estas deben formar parte del currículo de enseñanza de lenguas extranjeras (véase también Sinclair 1991). Nótese que en la bibliografía se utilizan diferentes términos – sin ser necesariamente sinónimos – para referirse al tipo o los tipos de expresiones en cuestión, como por ejemplo, *unidades fraseológicas, bloques semiprefabricados, chunks, expresiones multipalabra, fórmulas,* etc. (véase Wray 2002, 8–9).

Los resultados de estudios empíricos sobre la producción de los aprendices de L2 han ratificado la importancia de desarrollar el conocimiento de unidades multipalabra. De hecho, diversos autores concluyeron basándose en estudios de corpus de aprendices que el uso insuficiente o deficiente de estas expresiones es uno de los factores fundamentales que separan la producción de aprendices de nivel avanzado de la de hablantes nativos (Kjellmer 1991; Granger 1998; Howarth 1998). Al mismo tiempo, un repaso de la bibliografía relevante permite identificar varios argumentos a favor de la enseñanza de expresiones multilexémicas: se afirma que estas no solo contribuyen a una producción más parecida a la de hablantes nativos, sino también a una mayor fluidez; además, según los adscritos al llamado enfoque de aprendizaje basado en formulas (*formula-based approach*), estas expresiones tienen un papel clave en la adquisición del sistema lingüístico de la L2 (Durrant 2008, 40–57). A la hora de tratar de explicar por qué el aprendizaje de las expresiones multipalabra resulta problemático, Wray (2002) señala que el sistema cognitivo desarrollado de los aprendices adultos de L2, así como algunos factores sociales y comunivativos que difieren entre la adquisición de la primera lengua y el aprendizaje de L2, provocan que la atención de los aprendices se centre en las palabras simples, en vez de las unidades multilexémicas. Desde un enfoque diferente, Durrant (2008, 185), entre otros, afirma que el aspecto problemático de este tipo de expresiones se deriva de la cantidad insuficiente de *input*, que se debe al contacto más reducido con la lengua meta, y que afecta particularmente al aprendizaje de combinaciones menos frecuentes (véase también Henriksen 2013, 40–42).

Esta tesis trata de contribuir a la investigación relativa al conocimiento y uso de expresiones multilexémicas de aprendices de L2, centrándose en los aprendices de español como lengua extranjera (ELE). Asimismo, pretende explorar en qué manera las expresiones multipalabra deben ser representadas en las herramientas didácticas. Más en concreto, nos ocupamos del subtipo de unidades multipalabra constituido por las *colocaciones*. Utilizamos el término *colocación*, siguiendo la definición proporcionada por Mel'čuk (1998; 2012) dentro del enfoque de la *Lexicología explicativa y combinatoria*, para designar combinaciones binarias restringidas, como por ejemplo, *dar un paseo* o *fumador empedernido*. Dada la importancia de las colocaciones en el aprendizaje de L2, estas expresiones han constituido el objeto de numerosos estudios enmarcados en el ámbito de adquisición y enseñanza de segundas lenguas, como es el caso de estudios de corpus de aprendices (véase p. ej. Granger 1998; Howarth 1998; Nesselhauf 2005; Durrant y Schmitt 2009), estudios experimentales que tratan de abordar la competencia colocacional (véase Gyllstad 2007; Moreno Jaén 2009; Siyanova y Schmitt 2008), así como estudios que tienen como objetivo la evaluación y desarrollo de trabajos lexicográficos y materiales didácticos (véase Alonso Ramos 2008; Higueras García 2006; Komuro 2009; Lew y Radłowska 2010).

Objetivos principales

El objetivo de la presente tesis es examinar las necesidades de los aprendices de ELE en lo que respecta el desarrollo de su competencia colocacional. Esto implica estudiar el uso de colocaciones por parte de aprendices de ELE, así como los recursos didácticos existentes a–disposición del aprendiz. Tratamos de obtener a través de este estudio unos resultados que puedan ser aprovechados para el diseño de una nueva herramienta en línea destinada a aprendices de español.

Por consiguiente, la tesis trata las siguientes tres preguntas principales:

1. *¿Cuáles son las características de la producción colocacional de los aprendices de ELE?*

Dado que el presente trabajo pretende adoptar un enfoque del aprendizaje de colocaciones que tenga en cuenta las necesidades del aprendiz, nuestro punto de partida constituye necesariamente la descripción de la producción colocacional del grupo de aprendices que nos concierne. Es importante resaltar que los estudios existentes que han examinado las colocaciones utilizadas por aprendices de L2 han tratado casi exclusivamente el caso de aprendices de inglés. A pesar de que dentro del marco de ELE existen propuestas didácticas para la enseñanza de colocaciones (véase p. ej. Cavanillas 2008; Ferrando Aramo 2009; Higueras García 2006), estas se basan meramente en el supuesto que este tipo de expresiones multilexémicas generalmente resultan problemáticas, pero no están fundamentados en datos empíricos sólidos en lo que respecta el uso real de colocaciones por aprendices.

2. *¿Qué tipo de recursos enfocados en las colocaciones y dirigidos a aprendices de L2 existen, cómo son y con qué grado de éxito se utilizan?*

En la bibliografía se destaca tanto en el caso más general del aprendizaje de vocabulario (véase Nation 2001) como en el caso específico del aprendizaje de colocaciones (véase Hill et al.; Nesselhauf 2005; Woolard 2000) que los aprendices se deben dotar de estrategias que permiten el aprendizaje autónomo. Los diccionarios y los corpus lingüísticos constituyen dos tipos de recursos frecuentemente recomendados para ser utilizados de manera autónoma como apoyo a la producción de colocaciones. Existen, además, una serie de herramientas de corpus en línea que han sido desarrolladas a propósito para el uso por aprendices. El número de recursos disponibles para ELE es considerablemente más bajo que el de los existentes a disposición de aprendices de inglés, y, además, los estudios centrados en los aspectos de la usabilidad de este tipo de herramientas en general es escaso.

3. *¿Cómo debería ser una herramienta didáctica en línea enfocada en las colocaciones y dirigida a aprendices de ELE?*

Como ya ha sido mencionado, esta tesis trata de abordar el diseño de tal herramienta didáctica desde el punto de vista de las necesidades del aprendiz, lo que implica que se parte de los resultados de estudios empíricos relevantes. Nótese que tal y como se discute en la presente tesis, los recursos actualmente existentes no son suficientes, en el sentido de que en su diseño no se aprovecha el pleno potencial del soporte electrónico, no se observan todas las necesidades del usuario aprendiz a la hora de producir colocaciones, y además, las herramientas no se adecuan necesariamente a las destrezas del usuario potencial.

Estructura de la tesis y resultados

La presente tesis se divide en cinco capítulos principales enmarcados por un capítulo de introducción y uno de conclusiones. Los Capítulos 2 y 3 sirven para proporcionar un contexto teórico, además de repasar estudios previos relevantes, mientras que cada uno de los capítulos restantes está dedicado a tratar una de las preguntas presentadas arriba mediante la descripción de estudios originales.

El Capítulo 2 se ocupa de explorar el concepto de colocación dentro del ámbito de la lingüística. La primera parte del capítulo se centra en la noción de colocación, siguiendo su evolución a partir del surgimiento del término utilizado para referirse al fenómeno de co-ocurrencia léxica, y examinando las diferentes maneras en las que se ha definido desde el punto de vista de dos tradiciones teóricas principales: el enfoque basado en frecuencias y el enfoque fraseológico. Se hace una reflexión detallada con el objetivo de contrastar las ideas de diferentes autores en lo que respecta las características principales de las colocaciones con la definición del término proporcionada dentro del enfoque de la *Lexicología explicativa y combinatoria* (LEC, véase Mel'čuk et al. 1984; Mel'čuk et al. 1988; Mel'čuk et al. 1992; Mel'čuk et al. 1999), que hemos elegido como nuestro marco teórico. Es importante destacar, por tanto, que dentro de este enfoque las colocaciones se definen como combinaciones binarias léxicamente restringidas, compuestas por un elemento semántica y léxicamente autónomo, la *base*, que determina la selección del elemento restringido, el *colocativo*, para expresar un sentido dado.

La segunda parte de este capítulo trata dos aspectos principales de la descripción de colocaciones, considerados especialmente relevantes desde el punto de vista de la didáctica: el esquema sintáctico y el contenido semántico de las colocaciones. Ambos constituyen información clave en lo que se refiere a la presentación de estas expresiones en diccionarios combinatorios o diccionarios de colocaciones. Por tanto, repasamos los estudios de varios autores y obras lexicográficas que tratan sintácticos colocacionales, así como los diferentes enfoques sobre la descripción y clasificación de colocaciones según su significado. También introducimos el sistema de *funciones léxicas* (FFLL, Mel’čuk 1996), utilizado en la representación de las características sintácticas y semánticas de las colocaciones dentro del marco de la LEC.

El Capítulo 3 ofrece un panorama general de los enfoques teóricos y pedagógicos, así como de los estudios empíricos relacionados con el amplio campo de investigación que abarca la intersección de las expresiones multilexémicas y la enseñanza de L2. El capítulo trata tres aspectos principales. En primer lugar, se enumeran algunos de los argumentos que promueven la enseñanza de unidades multipalabra a aprendices de L2. Uno de estos argumentos se refiere a lo establecido por las teorías de adquisición que mantienen que los patrones lingüísticos se adquieren a través del análisis de secuencias multilexémicas memorizadas como simples bloques, ya que según algunos autores este proceso es relevante también en el caso del aprendizaje de lenguas extranjeras. Otros argumentos están fundamentados en modelos lingüísticos según los cuales una parte considerable del lenguaje utilizado por hablantes nativos está constituido por bloques pre-construidos, que, además, contribuyen a una producción más fluida.

En segundo lugar, el capítulo explora el uso y la competencia colocacional de aprendices de L2 mediante los resultados de estudios experimentales y estudios que usan corpus de aprendices. En líneas generales, los resultados empíricos muestran que el conocimiento de colocaciones de aprendices está en correlación con su nivel de competencia general, mientras su producción colocacional, que difiere tanto cuantitativamente como cualitativamente de la de hablantes nativos, se caracteriza por el sobreuso de expresiones frecuentes o favoritas, además de una cantidad considerable de combinaciones erróneas que son el resultado de transferencia de la L1.

El tercer y último aspecto tratado en el Capítulo 3 concierne la enseñanza y aprendizaje de colocaciones desde una perspectiva más cercana a la práctica didáctica. Consideramos, brevemente, algunas propuestas didácticas que – debido a la influencia de la investigación lingüística relevante – proponen la incorporación de las colocaciones en el currículo, tanto en el contexto de la enseñanza de inglés como lengua extranjera como en el ámbito de ELE. Un hilo común de las propuestas didácticas es el énfasis en la importancia de familiarizar a los alumnos tanto con la noción de colocación como con estrategias que permiten el aprendizaje autónomo y que pueden ser aplicadas fuera del aula. De acuerdo con esto, proporcionamos una descripción detallada de algunos de los recursos que los aprendices de L2 tienen a su disposición y que pueden utilizar de manera autónoma para mejorar su competencia colocacional y su uso de colocaciones: diccionarios de colocaciones, corpus lingüísticos y herramientas en línea.

El Capítulo 4 describe un estudio que analiza un corpus de aprendices de ELE para explorar su producción de colocaciones. Hemos considerado necesario llevar a cabo este tipo de estudio, debido a que, tal y como se observa en el Capítulo 3, los estudios empíricos existentes han tratado casi exclusivamente la competencia y uso de colocaciones de los aprendices de inglés. Por tanto, el estudio descrito en el Capítulo 4 tiene como finalidad contribuir a llenar el vacío ante la falta de este tipo de estudios, así como proporcionar una base empírica para proponer criterios de diseño en el caso de herramientas didácticas enfocadas en las colocaciones.

Nuestro estudio se basó en la anotación de una porción del *Corpus Escrito de Español como L2* (CEDEL2, Lozano 2009; Lozano y Mendikoetxea 2013) que consiste en 100 textos escritos por aprendices de ELE y 103 textos producidos por hablantes nativos, y en el consiguiente análisis de los datos. A la hora de comparar el uso de colocaciones de aprendices de ELE con el de hablantes nativos, podemos decir que nuestros resultados corroboran los de estudios previos sobre aprendices de inglés que señalaron que el uso de colocaciones de aprendices no se puede describir simplemente en términos de infrautilización, sino que debe estudiarse atendiendo a distintos fenómenos. Primero, teniendo en cuenta todas las colocaciones anotadas en los dos subcorpus, y sin considerar su patrón sintáctico, hemos observado que los aprendices de ELE producen una cantidad similar de colocaciones a la de los hablantes nativos. Segundo, como es esperable, nuestros datos muestran que los aprendices parecen utilizar un repertorio más reducido de combinaciones diferentes. Tercero, los datos referentes al uso de colocativos y, especialmente, al de colocativos verbales frecuentes, parecen estar en línea con observaciones de otros autores, según las cuales los aprendices tienden a abusar de un número reducido de elementos que creen poder utilizar con seguridad. Y, por último, los patrones de sobre- e infrautilización observados en el caso las combinaciones de los tipos VERBO+NOMBRE$_{\text{COMP}}$ y NOMBRE+ADJETIVO, respectivamente, plantean una cuestión en cuanto a la influencia de las estructuras más prominentes de la L1 sobre el uso de expresiones análogas en la L2. Al respecto, hemos notado que mientras en estudios previos se ha observado la infrautilización de las combinaciones del tipo VERBO+NOMBRE$_{\text{COMP}}$ en comparación con hablantes nativos (Altenberg y Granger 2001;

Howarth 1996; Laufer y Waldman 2011), nuestros datos que conciernen aprendices de español con L1 inglés demuestran una tendencia contraria.

En lo que concierne la tasa de error presentada por las colocaciones utilizadas por aprendices de ELE, según nuestros datos cerca de cuarta parte de las colocaciones anotadas en el subcorpus son erróneas. El uso de una tipología detallada de errores colocacionales a la hora de anotar el corpus nos ha permitido realizar un análisis más profundo de los mismos. Los resultados apuntan a que la mayoría de los errores colocacionales léxicos afectan el colocativo, lo que concuerda con el presupuesto implicado por la noción de colocación adoptada en esta tesis, es decir, con la hipótesis de que la selección del elemento léxicamente restringido de la combinación supone una mayor dificultad para los aprendices. Nuestros resultados relativos al origen del error son similares a los de otros estudios que muestran que la mayoría de errores léxicos constituyen el resultado de transferencia. Por último, a lo largo del proceso de anotación hemos encontrado algunos tipos de errores colocacionales que no se han tenido en cuenta en recursos o herramientas dirigidos a aprendices, como son, por ejemplo, los casos de errores gramaticales (**tomar sol*, **montar a bicicleta*), los errores léxicos que afectan la base (**lograr un gol* [objetivo]), y los errores léxicos que consisten en una palabra inexistente en la L2 (**ir de hiking, *misinterpretaciones* [malas interpretaciones]).

El Capítulo 5 presenta dos estudios experimentales. El primero de estos constituye un estudio de usabilidad que se centra en el *Diccionario de Colocaciones del Español* (DiCE, Alonso Ramos 2004), un diccionario en línea. Los resultados de este estudio sirven, por un lado, como indicación para mejorar el diseño de la interfaz del diccionario en cuestión, y, por otro lado, se consideran transferibles para el diseño de otras interfaces de diccionario o de otros tipos de herramientas de consulta. Hemos observado, por ejemplo, que los participantes preferían utilizar la opción de búsqueda por defecto del DiCE, que es similar a las consultas de diccionario habituales. Por un lado, creemos que esto sugiere que las interacciones de usuarios con diccionarios electrónicos se guían en gran medida por su experiencia previa en cuanto la consulta de diccionarios. Por otro lado, la falta de disposición del usuario para explorar los tipos de búsqueda alternativos y/o más avanzados nos lleva a concluir que probablemente sea más recomendable incorporar en una interfaz una opción de búsqueda universal, que permita descubrir el potencial completo de las búsquedas en el diccionario. Los resultados del experimento también resaltan el hecho de que los diccionarios deben evitar el uso extensivo de terminología lingüística, ya que los usuarios normalmente no están familiarizados con él, además de la importancia de la instrucción en el uso de diccionarios como parte del currículo de la enseñanza de L2.

La segunda parte del Capítulo 5 describe un experimento cuyo objetivo ha sido examinar la destreza de aprendices de ELE en la corrección de diferentes tipos de errores colocacionales con la ayuda de líneas de concordancias extraídas de corpus. Para ello, hemos seleccionado una serie de frases del corpus de aprendices analizado en el Capítulo 4 que contenían colocaciones erróneas y hemos creado un cuestionario donde cada una de estas frases va acompañada de ejemplos que representan el resultado de una búsqueda en corpus. Los resultados del experimento muestran un éxito general en la corrección de errores. Hemos observado que los participantes lograron corregir los errores léxicos con mayor grado de éxito que los errores gramaticales, probablemente debido a la falta de prominencia de estos últimos. Hemos encontrado, además, que los participantes consiguieron producir más reformulaciones correctas cuando se les ofrecieron concordancias con formato de frase completa que cuando recibieron concordancias en formato de n-gramas. En resumen, consideramos que los resultados del experimento son prometedores a la hora de evaluar la viabilidad de una herramienta basada en corpus y enfocada en el uso de colocaciones, dado que parecen demostrar que los aprendices de ELE son potencialmente capaces de interpretar datos lingüísticos extraídos de corpus y contrastarlos con su propia producción.

Finalmente, el Capítulo 6 describe el diseño de una herramienta de aprendizaje nueva centrada en colocaciones. La propuesta presentada se basa en la premisa de que el diseño de tal herramienta debe tener en cuenta las necesidades reales de aprendices, es decir, debe basarse en los resultados de estudios empíricos que tratan su conocimiento y uso de colocaciones, además de su destreza en lo que respecta el uso de herramientas y los resultados de aprendizaje derivados del mismo.

El recurso propuesto constituye principalmente una herramienta de consulta, e incorpora un diccionario de colocaciones y una herramienta de corpus, que se complementan entre sí a la hora de proporcionar información combinatoria. Al mismo tiempo, otras funcionalidades, como las opciones de crear un diccionario personal de colocaciones o la de generar ejercicios ofrecen una experiencia de aprendizaje personalizado. El capítulo describe las diferentes funcionalidades de la herramienta propuesta, con especial énfasis en la organización y presentación de la información combinatoria y de las posibilidades ofrecidas por la herramienta de consulta. La interfaz permitiría obtener información sobre la combinatoria de palabras individuales – de un modo parecido a los diccionarios de colocaciones –, encontrar las colocaciones de una palabra que expresan un sentido determinado, obtener información sobre el uso de una colocación concreta y verificar la corrección de las combinaciones utilizadas en un texto producido por el aprendiz.

En definitiva, esta tesis examina las necesidades de aprendices de ELE en lo que respecta el desarrollo de su competencia y uso colocacional, con el propósito de obtener resultados que sean de utilidad a la hora de diseñar una nueva herramienta didáctica centrada en las colocaciones. Los resultados del estudio de corpus de aprendices descrito en el Capítulo 4 permiten observar algunos de los aspectos problemáticos del uso de colocaciones, y proporcionan información en cuanto a los tipos de errores que afectan dichas combinaciones en la producción de aprendices. Al mismo tiempo, el estudio de usabilidad del DiCE y el experimento que trata la corrección de errores colocacionales con la ayuda de concordancias, descritos en el Capítulo 5, permiten evaluar la destreza de aprendices de ELE a la hora de interactuar con diferentes herramientas de manera autónoma. Como se demuestra en el Capítulo 6, los resultados de estos estudios pueden en efecto resultar provechosos a la hora de tomar decisiones en cuanto al diseño de una nueva herramienta didáctica.

Appendix G. Resumo da tese

Motivación e obxecto de estudo

Ao longo das últimas décadas, aumentou o interese con respecto ao vocabulario dentro do ámbito do ensino de segundas linguas (L2). Destacouse, ademais, que o coñecemento léxico non só se compón de palabras individuais, senón que tamén inclúe unha cantidade considerable de expresións multilexémicas. Este punto de vista reflíctese no traballo de Pawley e Syder (1983, 193) que afirman que, dentro das oracións gramaticais que se poden formular nunha lingua, só un subconxunto de oracións serían consideradas por falantes nativos como expresións normais, fronte ao resto que serían cualificadas como raras, non naturais ou estranxeirismos, ao non ter en conta que as palabras non se combinan libremente. En consecuencia, deféndese que un aprendiz non pode acadar un dominio da lingua semellante ao dun falante nativo sen ter un coñecemento suficiente das unidades multipalabra, polo tanto, estas deben formar parte do currículo do ensino de linguas estranxeiras (véxase tamén Sinclair 1991). Nótese que na bibliografía utilízanse diferentes termos – sen ser necesariamente sinónimos – para referirse ao tipo ou aos tipos de expresións en cuestión como, por exemplo, *unidades fraseolóxicas, bloques semiprefabricados, chunks, expresións multipalabra*, *fórmulas*, etc. (véxase Wray 2002, 8–9).

Os resultados de estudos empíricos sobre a produción dos aprendices de L2 ratificaron a importancia de desenvolver o coñecemento de unidades multipalabra. De feito, diversos autores concluíron, baseándose en estudos de corpus de aprendices, que o uso insuficiente ou deficiente destas expresións é un dos factores fundamentais que separan a produción de aprendices de nivel avanzado da de falantes nativos (Kjellmer 1991; Granger 1998; Howarth 1998). Ao mesmo tempo, un repaso da bibliografía relevante permite identificar varios argumentos en favor do ensino de expresións multilexémicas: afírmase que estas non só contribúen a unha produción máis parecida á de falantes nativos, senón tamén a unha maior fluidez; ademais, segundo os adscritos ao chamado enfoque de aprendizaxe baseado en fórmulas (*formula-based approach*), estas expresións teñen un papel clave na adquisición do sistema lingüístico da L2 (Durrant 2008, 40–57). Á hora de tratar de explicar por que a aprendizaxe das expresións multipalabra resulta problemática, Wray (2002) sinala que o sistema cognitivo desenvolvido dos aprendices adultos de L2, así como algúns factores sociais e comunicativos que difiren entre a adquisición da primeira lingua e a aprendizaxe de L2, provocan que a atención dos aprendices se centre nas palabras simples, no canto das unidades multilexémicas. Desde un enfoque diferente, Durrant (2008, 185), entre outros, afirma que o aspecto problemático deste tipo de expresións derívase da cantidade insuficiente de *input*, que se debe ao contacto máis reducido coa lingua meta, e que afecta particularmente á aprendizaxe de combinacións menos frecuentes (véxase tamén Henriksen 2013, 40–42).

Esta tese trata de contribuír á investigación relativa ao coñecemento e uso de expresións multilexémicas de aprendices de L2, centrándose nos aprendices de español como lingua estranxeira (ELE). Así mesmo, pretende explorar como as expresións multipalabra deben ser representadas nas ferramentas didácticas. Máis en concreto, ocupámonos do subtipo de unidades multipalabra constituído polas colocacións. Utilizamos o termo colocación, seguindo a definición proporcionada por Mel'čuk (1998; 2012) dentro do enfoque da *Lexicoloxía explicativa e combinatoria*, para designar combinacións binarias restrinxidas, como por exemplo, *dar un paseo* o *fumador empedernido*. Dada a importancia das colocacións na aprendizaxe de L2, estas expresións constituíron o obxecto de numerosos estudos enmarcados no ámbito de adquisición e ensino de linguas extranxeiras, como é o caso de estudos de corpus de aprendices (véxase p. ex. Granger 1998; Howarth 1998; Nesselhauf 2005; Durrant e Schmitt 2009), estudos experimentais que tratan de abordar a competencia colocacional (véxase Gyllstad 2007; Moreno Jaén 2009; Siyanova e Schmitt 2008), así como estudos que teñen como obxectivo a avaliación e o desenvolvemento de traballos lexicográficos e materiais didácticos (véxase Alonso Ramos 2008; Higueras García 2006; Komuro 2009; Lew e Radłowska 2010).

Obxectivos principais

O obxectivo da presente tese é examinar as necesidades dos aprendices de ELE no referente ao desenvolvemento da súa competencia colocacional. Isto implica estudar o uso de colocacións por parte de aprendices de ELE, así como o dos recursos didácticos existentes á disposición do aprendiz. Tratamos de obter a través destes estudos uns resultados que poidan ser aproveitados para o deseño dunha nova ferramenta en liña destinada a aprendices de español.

Por conseguinte, a tese trata as seguintes tres preguntas principais:

1. *Cales son as características da produción colocacional dos aprendices de ELE*?

Dado que o presente traballo pretende adoptar un enfoque da aprendizaxe de colocacións que teña en conta as necesidades do aprendiz, o noso punto de partida constitúe necesariamente a descrición da produción colocacional do grupo de aprendices que nos concirne. É importante resaltar que os estudos existentes que examinaron as colocacións utilizadas por aprendices de L2 trataron case exclusivamente o caso de aprendices de inglés. A pesar de que dentro do marco de ELE existen propostas didácticas para o ensino de colocacións (véxase p. ex. Cavanillas 2008; Ferrando Aramo 2009; Higueras García 2006), estas baséanse meramente no suposto de que este tipo de expresións multilexémicas xeralmente resultan problemáticas, pero non están fundamentadas en datos empíricos sólidos no referente ao uso real de colocacións por aprendices.

2. *Que tipo de recursos enfocados nas colocacións e dirixidos a aprendices de L2 existen, como son e con que grao de éxito se utilizan?*

Na bibliografía destácase tanto no caso máis xeral da aprendizaxe de vocabulario (véxase Nation 2001) como no caso específico da aprendizaxe de colocacións (véxase Hill *et al*.; Nesselhauf 2005; Woolard 2000) que os aprendices deben dotarse de estratexias que permitan a aprendizaxe autónoma. Os dicionarios e os corpus lingüísticos constitúen dous tipos de recursos frecuentemente recomendados para ser utilizados de xeito autónomo como apoio á produción de colocacións. Alén disto, existe unha serie de ferramentas de corpus en liña que foron desenvolvidas a propósito para os aprendices usaren. O número de recursos dispoñibles para ELE é considerablemente máis baixo que o dos existentes a disposición de aprendices de inglés, e, ademais, os estudos centrados nos aspectos da usabilidade deste tipo de ferramentas, en xeral, é escaso.

3. *Como debería ser unha ferramenta didáctica en liña enfocada nas colocacións e dirixida a aprendices de ELE?*

Como xa foi mencionado, esta tese trata de abordar o deseño de tal ferramenta didáctica desde o punto de vista das necesidades do aprendiz, feito que implica que se parta dos resultados de estudos empíricos relevantes. Nótese que tal e como se discute na presente tese, os recursos actualmente existentes non son suficientes, no sentido de que no seu deseño non se aproveita o pleno potencial do soporte electrónico, non se observan todas as necesidades do usuario aprendiz á hora de producir colocacións e as ferramentas non se adecúan necesariamente ás destrezas do usuario potencial.

Estrutura da tese e resultados

A presente tese divídese en cinco capítulos principais enmarcados por un capítulo de introdución e outro de conclusións. Os Capítulos 2 e 3 proporcionan un contexto teórico, ademais de repasar estudos previos relevantes; mentres cada un dos capítulos restantes está dedicado a tratar unha das preguntas presentadas arriba mediante a descrición de estudos orixinais.

O Capítulo 2 explora o concepto de colocación no ámbito da lingüística. A primeira parte do capítulo céntrase na noción de colocación, segue despois a súa evolución a partir do xurdimento do termo utilizado para referirse ao fenómeno de co-ocorrencia léxica, examinando os diferentes xeitos en que se definiu desde o punto de vista de dúas tradicións teóricas principais: o enfoque baseado en frecuencias e o enfoque fraseolóxico. Faise unha reflexión detallada co obxectivo de contrastar as ideas de diferentes autores no tocante ás características principais das colocacións coa definición do termo proporcionada polo enfoque da Lexicoloxía explicativa e combinatoria (LEC, véxase Mel'čuk et al. 1984; Mel'čuk et al. 1988; Mel'čuk et al. 1992; Mel'čuk et al. 1999), que escollemos como o noso marco teórico. Cómpre destacar, polo tanto, que dentro deste enfoque as colocacións defínense como combinacións binarias lexicamente restrinxidas, compostas por un elemento semántica e lexicamente autónomo, a base, que determina a selección do elemento restrinxido, o colocativo, para expresar un sentido dado.

A segunda parte deste capítulo trata dos aspectos principais da descrición de colocacións, considerados especialmente relevantes desde o punto de vista da didáctica: o esquema sintáctico e o contido semántico das colocacións. Ambos os dous constitúen información clave no que se refire á presentación destas expresións en dicionarios combinatorios ou dicionarios de colocacións. Polo tanto, repasamos os estudos de varios autores e obras lexicográficas que tratan os posibles esquemas sintácticos colocacionais, así como os diferentes enfoques sobre a descrición e clasificación de colocacións segundo o seu significado. Tamén introducimos o sistema de *funcións léxicas* (FFLL, Mel'čuk

1996), utilizado na representación das características sintácticas e semánticas das colocacións dentro do marco da LEC.

O Capítulo 3 ofrece un panorama xeral dos enfoques teóricos e pedagóxicos, así como dos estudos empíricos relacionados co amplo campo de investigación que abrangue a intersección das expresións multilexémicas e o ensino de L2. O capítulo trata tres aspectos principais. En primeiro lugar, enuméranse algúns dos argumentos que promoven o ensino de unidades multipalabra para aprendices de L2. Un destes argumentos refírese ao establecido polas teorías de adquisición que manteñen que os patróns lingüísticos se adquiren a través da análise de secuencias multilexémicas memorizadas como simples bloques, xa que segundo algúns autores este proceso é relevante tamén no caso da aprendizaxe de linguas estranxeiras. Outros argumentos están fundamentados en modelos lingüísticos segundo os cales unha parte considerable da linguaxe utilizada por falantes nativos está constituída por bloques pre-construídos, que, ademais, contribúen a unha produción máis fluída.

En segundo lugar, o capítulo explora o uso e a competencia colocacional de aprendices de L2 mediante os resultados de estudos experimentais e estudos que usan corpus de aprendices. En liñas xerais, os resultados empíricos mostran que o coñecemento de colocacións de aprendices está en correlación co seu nivel de competencia xeral, mentres a súa produción colocacional, que difire tanto cuantitativa como cualitativamente da de falantes nativos, caracterízase polo sobreuso de expresións frecuentes ou favoritas, ademais dunha cantidade considerable de combinacións erróneas orixinadas pola transferencia da L1.

O terceiro e último aspecto tratado no Capítulo 3 concirne ao ensino e aprendizaxe de colocacións desde unha perspectiva máis achegada á práctica didáctica. Consideramos, brevemente, algunhas propostas didácticas que –debido á influencia da investigación lingüística relevante– propoñen a incorporación das colocacións no currículo, tanto no contexto do ensino de inglés como lingua estranxeira como no ámbito de ELE. Un fío común das propostas didácticas é a énfase na importancia de familiarizar os alumnos tanto coa noción de colocación coma con estratexias que permitan a aprendizaxe autónoma e que poidan ser aplicadas fóra da aula. De acordo con isto, proporcionamos unha descrición detallada dalgúns dos recursos que os aprendices de L2 teñen á súa disposición e que poden utilizar de xeito autónomo para melloraren a súa competencia colocacional e o seu uso de colocacións: dicionarios de colocacións, corpus lingüísticos e ferramentas en liña.

O Capítulo 4 describe un estudo que analiza un corpus de aprendices de ELE para explorar a súa produción de colocacións. Consideramos necesario levar a cabo este tipo de estudo, debido a que, tal e como se observa no Capítulo 3, os estudos empíricos existentes trataron case exclusivamente a competencia e uso de colocacións dos aprendices de inglés. Polo tanto, o estudo descrito no Capítulo 4 ten como finalidade contribuír a encher o baleiro deste tipo de estudos, así como proporcionar unha base empírica propoñendo criterios de deseño no caso de ferramentas didácticas enfocadas nas colocacións.

O noso estudo baseouse na anotación dunha porción do *Corpus Escrito de Español como L2* (CEDEL2, Lozano 2009; Lozano e Mendikoetxea 2013) que consiste en 100 textos escritos por aprendices de ELE e 103 textos producidos por falantes nativos, e na conseguinte análise dos datos. Despois de comparar o uso de colocacións de aprendices de ELE co de falantes nativos, podemos dicir que os nosos resultados corroboran os de estudos previos sobre aprendices de inglés que sinalaron que o uso de colocacións de aprendices non se pode describir simplemente en termos de infrautilización, senón que debe estudarse atendendo a distintos fenómenos. Primeiro, tendo en conta todas as colocacións anotadas nos dous subcorpus, e sen considerar o seu patrón sintáctico, observamos que os aprendices de ELE producen unha cantidade de colocacións similar á dos falantes nativos. Segundo, como é esperable, os nosos datos mostran que os aprendices parecen utilizar un repertorio máis reducido de combinacións diferentes. Terceiro, os datos alusivos ao uso de colocativos e, especialmente, ao de colocativos verbais frecuentes, parecen estar en liña con observacións doutros autores, segundo as cales os aprendices tenden a abusar dun número reducido de elementos que cren poder utilizar con seguridade. E, por último, atendendo aos patróns de sobre- e infrautilización observados no caso das combinacións dos tipos VERBO+NOME$_{\text{COMP}}$ e NOME+ADXECTIVO, respectivamente, xorde unha cuestión atinente á influencia das estruturas máis prominentes da L1 sobre o uso de expresións análogas na L2. Ao respecto, notamos que mentres en estudos previos se observou a infrautilización das combinacións do tipo VERBO+NOME$_{\text{COMP}}$ en comparación con falantes nativos (Altenberg e Granger 2001; Howarth 1996; Laufer e Waldman 2011), os nosos datos relativos a aprendices de español con L1 inglés demostran unha tendencia contraria.

No concernente á taxa de erro presentada polas colocacións utilizadas por aprendices de ELE, segundo os nosos datos preto da cuarta parte das colocacións anotadas no subcorpus son erróneas. O uso dunha tipoloxía detallada de erros colocacionais á hora de anotar o corpus permitiunos realizar unha análise máis profunda deles. Os resultados apuntan a que a maioría dos erros colocacionais léxicos afectan ao colocativo, feito que concorda co presuposto implicado pola noción de colocación adoptada nesta tese, é dicir, coa hipótese de que a selección do elemento lexicamente restrinxido da combinación supón unha maior dificultade para os aprendices. En canto á orixe do erro, os nosos resultados son similares ao dos outros estudos que mostran que a maioría de erros léxicos son produto do resultado de transferencia. Por último, ao longo do proceso de anotación encontramos algúns tipos de erros colocacionais que non se tiveron en conta en recursos ou ferramentas dirixidos a aprendices, como son, por exemplo, os casos de erros gramaticais (**tomar sol, *montar a bicicleta*), os erros léxicos que afectan a base (**lograr un gol* [obxectivo]) e os erros léxicos que consisten nunha palabra inexistente na L2 (**ir de hiking, *misinterpretaciones* [malas interpretacións]).

O Capítulo 5 presenta dous estudos experimentais. O primeiro deles constitúe un estudo de usabilidade centrado no *Diccionario de Colocaciones del Español* (DiCE, Alonso Ramos 2004), un dicionario en liña. Os resultados deste estudo serven, por un lado, como indicación para mellorar o deseño da interface do dicionario en cuestión e, por outro lado, considéranse transferibles para o deseño doutras interfaces de dicionario ou doutros tipos de ferramentas de consulta. Observamos, por exemplo, que os participantes preferían utilizar a opción de procura por defecto do DiCE, similar ás consultas de dicionario habituais. Por unha banda, cremos que isto suxire que as interaccións de usuarios con dicionarios electrónicos guíanse en gran medida pola súa experiencia previa en canto á consulta de dicionarios. Por outra banda, a falta de disposición do usuario para explorar os tipos de procura alternativos e/ou máis avanzados lévanos a concluír que probablemente sexa máis recomendable incorporar nunha interface unha opción de procura universal, que permita descubrir o potencial completo das procuras no dicionario. Os resultados do experimento tamén resaltan o feito de que os dicionarios deben evitar o uso extensivo de terminoloxía lingüística, xa que os usuarios normalmente non están familiarizados con el, ademais da importancia da instrución no uso de dicionarios como parte do currículo do ensino de L2.

A segunda parte do Capítulo 5 describe un experimento cuxo obxectivo foi examinar a destreza de aprendices de ELE na corrección de diferentes tipos de erros colocacionais coa axuda de liñas de concordancias extraídas de corpus. Para iso, seleccionamos unha serie de frases do corpus de aprendices analizada no Capítulo 4 que contiñan colocacións erróneas. Logo creamos un cuestionario con cada unha desas frases, acompañada de exemplos que representan o resultado dunha procura en corpus. Os resultados do experimento mostran un éxito xeral na corrección de erros. Observamos que os participantes lograron corrixir os erros léxicos con maior grao de éxito que os erros gramaticais, probablemente debido á falta de prominencia destes últimos. Descubrimos, ademais, que os participantes conseguiron producir máis reformulacións correctas cando se lles ofreceron concordancias con formato de frase completa que cando recibiron concordancias en formato de n-gramas. En resumo, consideramos que os resultados do experimento son prometedores á hora de avaliar a viabilidade dunha ferramenta baseada en corpus e enfocada no uso de colocacións, dado que parecen demostrar que os aprendices de ELE son potencialmente capaces de interpretaren datos lingüísticos extraídos de corpus e contrastalos coa produción de seu.

Finalmente, o Capítulo 6 describe o deseño dunha ferramenta de aprendizaxe nova centrada en colocacións. A proposta presentada baséase na premisa de que o deseño de tal ferramenta debe ter en conta as necesidades reais de aprendices, é dicir, debe basearse nos resultados de estudos empíricos que tratan o seu coñecemento e uso de colocacións, ademais da súa destreza no uso de ferramentas e nos resultados de aprendizaxe derivados del.

O recurso proposto constitúe principalmente unha ferramenta de consulta, e incorpora un dicionario de colocacións e unha ferramenta de corpus, que se complementan entre si para proporcionar información combinatoria. Asemade, outras funcionalidades, como as opcións de crear un dicionario persoal de colocacións ou a de xerar exercicios ofrecen unha experiencia de aprendizaxe personalizada. O capítulo describe as diferentes funcionalidades da ferramenta proposta, con especial énfase na organización e presentación da información combinatoria e das posibilidades ofrecidas pola ferramenta de consulta. A interface permitiría obter información sobre a combinatoria de palabras individuais – dun modo semellante aos dicionarios de colocacións –, atopar as colocacións dunha palabra que expresen un sentido determinado, obter información sobre o uso dunha colocación concreta e verificar a corrección das combinacións utilizadas nun texto producido polo aprendiz.

En definitiva, esta tese examina as necesidades de aprendices de ELE respecto ao desenvolvemento da súa competencia e uso colocacional, co propósito de obter resultados que sexan de utilidade á hora de deseñar unha nova ferramenta didáctica centrada nas colocacións. Os resultados do estudo de corpus de aprendices descrito no Capítulo 4 permiten observar algúns dos aspectos problemáticos do uso de colocacións, e proporcionan información referente aos tipos de erros que afectan a esas combinacións na produción de aprendices. Ao mesmo tempo, o estudo de usabilidade do DiCE e o experimento que trata a corrección de erros colocacionais coa axuda de concordancias, descritos no Capítulo 5, permiten avaliar a destreza de aprendices de ELE no momento de interactuar con diferentes ferramentas de xeito autónomo. Como se demostra no Capítulo 6, os resultados destes estudos poden, en efecto, resultar proveitosos á hora de tomar decisións no deseño dunha nova ferramenta didáctica.