Recibido / Received: 10/11/2019
Aceptado / Accepted: 28/01/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2020.ne6.9

Para citar este artículo / To cite this article:
Rojas Díaz, José Luis. (2020) "From head to toe: A lexical, semantic, and morphosyntactic study of idioms in phraseological dictionaries in English and Spanish." In: Mogorrón Huerta, Pedro (ed.) 2020. *Análisis multidisciplinar del fenómeno de la variación fraseológica en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting. MonTI Special Issue* 6, pp. 287-326.

"FROM HEAD TO TOE": A LEXICAL, SEMANTIC, AND MORPHOSYNTACTIC STUDY OF IDIOMS IN PHRASEOLOGICAL DICTIONARIES IN ENGLISH AND SPANISH

José Luis Rojas Díaz
jose.diaz@nhh.no
NHH Norwegian School of Economics

Abstract

In recent times, interest in the study of phraseology in general and specialized lexicographic resources has increased (Castillo Carballo 2006, Aguado de Cea 2007, Mellado 2008, Buendía Castro & Faber 2015). However, to date, a lack of knowledge related to the characterization and indexation of phraseological units (PUs) in lexicographic resources remains. That issue is addressed here through an analysis of phraseological units in the entries of two phraseological dictionaries, one in Spanish, and one in English: the *Diccionario fraseológico documentado del español actual* (Seco, Andrés & Ramos 2004) and the *Collins COBUILD Dictionary of Idioms* (Sinclair & Moon 1997). To perform this analysis, two databases containing 21,045 entries extracted from the two dictionaries mentioned above were compiled. The databases were tagged syntactically and semantically in order to extract 816 morphosyntactic patterns, 2,655 combinations of semantic categories (Semantic patterns) and a series of lexical and lexicographic information about indexation of PUs in dictionaries.

Keywords: Phraseology; Dictionaries; Lexicography; Semantic patterns; Morphosyntactic patterns.

Resumen

En los últimos años ha aumentado el interés en el estudio de la fraseología en lengua general y de especialidad en recursos lexicográficos (Castillo Carballo 2006, Aguado de Cea 2007, Mellado 2008, Buendía Castro & Faber 2015). Sin embargo y a la fecha, aún existen algunos vacíos en cuanto a la caracterización e indización de unidades fraseológicas (UFs) en recursos lexicográficos. Esta problemática se aborda en el presente artículo por medio del análisis de unidades fraseológicas en las entradas de dos diccionarios (uno en español y otro en inglés): el *Diccionario fraseológico documentado del español actual* (Seco, Andrés & Ramos 2004) y el *Collins COBUILD Dictionary of Idioms* (Sinclair & Moon 1997). Para llevar a cabo este análisis, se compilaron dos bases de datos con 21 045 entradas extraídas de los diccionarios antes mencionados. Las bases de datos fueron etiquetadas sintáctica y semánticamente para extraer 816 patrones morfosintácticos, 2 655 combinaciones semánticas (patrones semánticos) y una serie de datos léxicos y lexicográficos sobre la indización de UFs en diccionarios.

Palabras clave: Fraseología; Diccionarios; Lexicografía; Patrones semánticos; Patrones morfosintácticos.

1. Introduction

In recent times, the study of phraseology in general language and specialized language lexicographic resources has gained particular interest (e.g. dictionaries and databases) (Castillo Carballo 2006: 8, Aguado de Cea 2007: 184-185, Mellado 2008, Buendía Castro & Faber 2015: 161). However, more in-depth knowledge is needed about the characterization and indexation of phraseological units.

This article will shed light on how PUs are indexed in dictionaries as well as the lexicographic, lexical, semantic, and morphosyntactic characteristics of the selected PUs. The analyses and the article are structured as follows: firstly, a summary of the different concepts regarding phraseology in the Spanish and English traditions is presented. Secondly, the lexicographic description of the dictionaries used for the compilation of the database is introduced. Thirdly, the results of the lexical, semantic and morphosyntactic analyses are presented. Lastly, the final section is devoted to the most salient conclusions reached, and to a practical lexicographic proposal for the indexation of PUs in lexicographic resources.

1.1. Phraseology: denomination and definition

According to García-Page (2008), phraseology should be defined in terms of its object of study. Thus, the question to ask would be "what is the object of study of phraseology?" (2008: 7). However, answering this question entails a problem, namely: the extensive number of denominations and definitions used to determine the object of study of phraseology. Bushnaq (2015: 173) states that the terms 'phraseologism', 'phraseme', 'phraseological expression', 'phraseological unit', 'idiomatic expression', and 'idiom' are used in English to describe an expression the meaning of which cannot be deduced from the individual meaning of its constituent words. Although the definition given by Bushnaq is correct, it is still vague and corresponds to the classical definition of phraseology. Among theoreticians, it is possible to find the most diverse taxonomies to categorize expressions according to their compositionality/idiomaticity, functional categories, and fixation in language, among other features. Ruiz Gurrillo (2001: 44) and Cowie (2001: 7) present a summary of those categories for some phraseology research in both Spanish and English. However, three categories are common to almost all taxonomies. Those categories include: (i) expressions that behave as sentences (proverbs/sayings), (ii) other expressions in which one of the constituents is not idiomatic (collocations), and (iii) other that are fully idiomatized (idioms).

The approach to the study of phraseological units in Spanish and English is considered to have major differences. On the one hand, the Spanish tradition tends to be taxonomic in nature, having two fundamental notions on the study of phraseology: a narrow one —in which only idioms are considered to be PUs—, and a wide one —where not only idioms, but also sayings, proverbs, collocations, among others are considered to be PUs—. On the other hand, the English tradition is more flexible —similar to the wide notion of Spanish phraseology—, and it includes many subsets of phrases that would not be considered as *phraseological unit*s by some theoreticians in Spanish. Particularly, the most restricted subset of units in both languages will be used in this paper, i.e. *locuciones* in Spanish and *idioms* in English, and they will be referred to as *phraseological units* in an attempt to use a denomination that encompasses the characteristics of both subsets in both languages. When looking up the definitions of *locución* and *idiom* in general dictionaries in

Spanish and in English, it becomes evident that those definitions evidently differ. Thus, on the one hand, the *Diccionario de la lengua española* (DLE [online]) defines *locución* in its third sense as a “group of words that function as a single lexical unit with a unitary meaning and certain degree of formal fixation” (locución, n.d., author’s translation). On the other hand, the *Cambridge Dictionary* (online) defines *idiom* as “a group of words in a fixed order that have a particular meaning that is different from the meanings of each word on its own” (idiom, n.d.).

In Spanish, when resorting to the literature on phraseology studies, one of the most accepted definitions of *locución* is given by Casares (1950), who states that a *locución* is: “a stable combination of two or more terms that function as an element in a sentence and whose unitary meaning cannot be simply justified as the sum of the usual meanings of its components” (1950: 170, author’s translation). As a further elaboration to the conception of Casares, Ruiz Gurillo (1997) says that a ‘phraseological unit’ is “a fixed combination of words that presents a certain level of fixation, and eventually, idiomaticity” (1997: 14). Likewise, in English, Moon (1998) states that the definition of idiom is ambiguous due to its different uses. Nonetheless, the author also asserts that the most restrictive definition of idioms could be “a particular kind of unit: one that is fixed and semantically opaque or metaphorical, or, traditionally, not the sum of its parts” (1998: 4). Similarly, Mel’čuk (2012) proposes a definition of ‘pure idiom’ in the following terms: “an idiom AB is a full idiom if its meaning does not include the meaning of any of its lexical components: ‘AB’ ⊅ ‘A’ and ‘AB’ ⊅ ‘B’” (2012: 37). This last definition put forward by Mel’čuk, will be the one applied to ‘phraseological units’ in this paper. In the next section, the Spanish and English phraseology traditions will be discussed in more detail.

1.2. Spanish and English theoretical traditions on phraseology

Norrick (2007) states that there are two different traditions related to the study of phraseology in English: the British tradition, and the American one. He also suggests that both traditions were originally driven by either anthropological or literary approaches (2007: 615). For the British tradition, Norrick proposes three stages in the study of phraseology. The first one is based on the "list of the irregularities in a language" written by Bloomfield (1933). The second one is the study conducted by Hockett (1958), where he grouped phraseological units in a category called idioms. The third one was the grammar written for those units by Householder (1959). The distinction between idioms and collocations was made, among others, by Firth (1957), and later by Sinclair & Moon (1989, 1997).

According to Norrick (2007: 616), the American tradition started with the criticism Wallace Chafe made of Noam Chomsky's compositionality concept. Chomsky argued that the lexicon is a "simple and unordered list of all lexical formatives" (Chomsky 1965: 84) which should include the idioms. Three years later, Chafe (1968) showed that the concept of idiomaticity, one of the characteristics of phraseological units, is totally opposed to the compositionality criterion of Chomskyan linguistic theory. Table 1 presents a synthesis of phraseological denominations in the English language according to Cowie (2001).

Table 1. Denominations of phraseology used in English by different authors according to Cowie (2001: 7)

Author	General category	Opaque, invariable unit	Partially motivated unit	Phraseological bound unit
Vinogradov	Phraseological unit	Phraseological fusion	Phraseological unity	Phraseological combination
Amosova	Phraseological unit	Idiom	Idiom (not differentiated)	Phraseme / Phraseoloid
Cowie	Composite	Pure idiom	Figurative Idiom	Restricted collocation
Mel'cuk	Semantic Phraseme	Idiom	Idiom (not differentiated)	Collocation
Gläser	Nomination	Idiom	Idiom (not differentiated)	Restricted collocation
Howarth	Composite unit	Pure Idiom	Figurative Idiom	Restricted collocation

In Spanish, authors such as Casares (1950), Zuluaga (1980), Carneado & Trista (1985), Corpas Pastor (1996), Ruiz Gurillo (2001), and García-Page

(2008) are among the most quoted ones in phraseology studies. Nevertheless, the denominations of idiomatized units proposed by those authors differ greatly, as shown in Table 2.

Table 2. Denominations of PUs proposed by the most representative authors related to general phraseology in Spanish

Author	Denomination	Definition
Julio Casares (1950)	*Locuciones*	Wide
	Frases hechas	
	Refranes	
	Modismos	
Alberto Zuluaga (1980)	*Locuciones*	
	Enunciados	
Zoila Carneado & Antonia Tristá (1985)	*Unidad fraseológica (fraseologismo) [verbal, reflexivo, propositivo, participial, conjuntivo, pronominal, nominal, adjetival, adverbial]*	
Gloria Corpas Pastor (1996)	*Unidad fraseológica [Colocación, Locución, Enunciado Fraseológico]*	
Leonor Ruiz Gurillo (2001)	*Locuciones [nominal, adjetival, verbal, adverbial, marcadora, propositiva, clausal]*	Narrow
Mario García-Page (2008)	*Locuciones [nominales, adjetivales, adverbiales, propositivas, conjuntivas, verbales, oracionales]*	

The Spanish tradition of the study of phraseology includes two basic conceptions: the wide one and the narrow one. The wide conception could include everything from proverbs or collocations (depending on the author) to idioms. The narrow conception focuses only on *locuciones* (idioms), as evidenced by works such as those by Carneado & Trista (1985: 68), Ruiz Gurillo (1998: 12), Rakotojoelimaria (2004: 25), Sosiński (2006: 23), Školníková (2010: 7), and López (2012: 57).

Although both the English and the Spanish traditions have denominations for each kind of PU, and authors have undoubtedly developed complex taxonomies to classify them, there are some aspects related to semantics and pragmatics that have not yet been addressed. For instance, literature on phraseology lacks information related to PUs' semantic patterns, or to the way in which their two semantic macro-components —the figurative and the mental image (Molina Plaza 2005: 176)— change from one language to another. This limitation is due to the lack of linguistic information —related to the composition of the PUs, their meaning, and how they are used in a communicative context— which can only be obtained through descriptive studies.

1.3. Phraseology and lexicography: a shared-ground proposal

In order to deal with phraseological units in dictionaries, it is necessary to talk about lexicography in general, and lexicographic resources (i.e. dictionaries, glossaries, databases) in particular. Lexicography is considered as an applied discipline related to linguistics. According to Sinclair (1984):

> "It is clearly an applied science or craft, rather than a pure one. That is to say, it relies for a theoretical framework on external disciplines. I know this is a contentious point and that this paper is not the proper forum for its debate, but the shape proposed for lexicography as an academic subject depends on the attitude taken to this issue. There is, for example, no subject heading 'Lexicography theory' in my syllabus because I have nothing to put there; on the other hand there is substantial input from IT and LINGUISTICS because I believe that the relevant theory is to be found in these areas or via these areas". (1984: 6-7).

According to Moon (2009), Sinclair showed that lexicography does not have a theoretical background due to its applied nature, but at the same time, she recognizes that the methodology Sinclair developed for the COBUILD project was based on principles that could be applied to lexicography in general, one of them being the use of corpus linguistics for the creation of the dictionary.

On lexicographers and lexicography, Atkins & Rundell (2008) state that "by the nature of the work they do, lexicographers are applied linguists", and although these authors think "a grounding in linguistic theory is not a prerequisite", they also believe that "there are certain basic linguistic concepts which are invaluable in preparing people to analyze data and to produce concise, accurate dictionary entries" (2008: 130). In turn, regarding phraseology and lexicography, Leroyer (2006) states that the relationship between these two disciplines should be considered a "scientific marriage" since they have been related for a long time. According to him, more than 1,700 reference entries can be found in the EURALEX site concerning both phraseology and lexicography (2006: 183). Leroyer also suggests that there are two ways to look at the relationship between these two disciplines: firstly, the treatment of phraseology by lexicographers and, secondly, the phraseological studies of linguists drawing recommendations on how to deal with phraseology in dictionaries (2006: 183). Furthermore, Paquot (2015) draws attention to several problems related to the phraseological information (related to collocations) that dictionaries provide. Among her findings, Paquot found a systematic lack of consistency in dictionary entries (2015: 5-6). This problem is also tackled by Moon (2008). She explains that lexicographic resources struggle with providing the description of phraseological units that meet the requirements of phraseological theories, and with the evidences of occurrence of those units in real texts. She further states that dictionaries must provide information about how idioms behave in context (2008: 314).

The study of the inclusion of phraseology in dictionaries has not only been of interest to linguists in English. It is also possible to find a number of articles related to the study of phraseology and dictionaries in Spanish. For instance, the papers in two books edited by Alonso (*Diccionarios y fraseología,* 2006) and Mellado Blanco (*Colocaciones y fraseología en los diccionarios,* 2008). On the one hand, included in the book edited by Alonso (2006), the study by González (2006) addresses how collocations and idioms are registered in the DRAE (Spanish Royal Academy's Dictionary of the Spanish Language). This study made by González arrives at the conclusion that the selection criteria for the inclusion of collocations follow the classification system developed by Corpas Pastor (1996), while idioms are categorized using the taxonomy proposed by Casares (1950). On the other hand, also included in the book edited by Alonso, the work by Penadés (2006: 252-253) discusses issues related to the marking of phraseological units in dictionaries.

In the book edited by Mellado Blanco (2008), Ortega Ojeda & González Aguilar present the marks used in two general language dictionaries in Spanish, and they conclude that the marking in both dictionaries is inaccurate. The same holds true for the criteria that lexicographers used to classify and mark PUs in the dictionaries studied (González Aguilar 2008: 244).

All these studies show the tendency for marking and indexation in dictionaries to be incomplete, inconsistent, or inaccurate to some extent. In addition, Buendía Castro & Faber (2015) state that phraseological units have begun to be indexed more frequently in dictionaries in recent years (2015: 161). However, this does not mean that a systematic methodology is followed for the indexation or lemmatization of phraseological entries —this includes, for example, the criteria for choosing a certain word form as the headword of the PU—. One possible explanation for this problem is that although many studies and theoretical-methodological reflections have been proposed on how to deal with phraseological units in dictionaries, the conclusions provided by such works do not seem to be taken into account in the lexicographic practice. For instance, the introduction or guidelines of dictionaries should include information regarding the marking and indexation of phraseological units (Santamaría Pérez 2003: 1045), but that is not always the case.

As shown above (section 1.1), it is possible to find concepts in Spanish and English that are applicable to all the PUs suitable to be indexed in a monolingual or a bilingual dictionary. However, these criteria must be synthetized and shared among experts and publishing houses in an attempt to reach a consensus in aspects such as taxonomy, selection criteria, and marking, as it has been done before in lexicographic manuals regarding monolexical entries. As for the representation and indexation of PUs, Heid (2008) states that many current projects and initiatives involving Natural Language Processing are taking place in relation to the development of standards for PUs. Nevertheless, problems regarding the automatic identification, extraction and productivity of PUs "are far from being solved" (2008: 349-350).

The "quantification of the phenomenon" and the succeeding recording of PUs (Heid 2008: 349-350) is one of the several challenges that lexicography faces regarding phraseology. On this matter, Jackendoff (1997) observes that "there are vast numbers of such memorized fixed expressions; these extremely crude estimates suggest that their number is of about the same order of magnitude as the single words of the vocabulary" (1997: 156). Jackendoff's claim is in turn quoted by Tschichold (2008) to add that the recording process of such amount of PUs in a language will always be incomplete (2008: 366). Heid (2008) identifies the need for more morphosyntactic and semantic annotated resources as well as research on this aspect of phraseology (2008: 354).

Nevertheless, he also points out that a possible solution for the identification of PUs could be reached by means of "distributional semantics" meaning that "items with similar contexts share meaning components" (Heid 2008: 353). A similar approach is used in this article (see sections 3.2 & 3.3) through the use of semantic annotation for the extraction of semantic patterns that could be used as criteria for the identification and extraction of PUs.

2. Data, Tools, and Methods

For the analysis intended here, two dictionaries were used: the *Diccionario fraseológico documentado del español actual* (henceforth DFDEA) (Seco, Andrés & Ramos 2004) and *The Collins COLBUILD dictionary of idioms* (henceforth CCDOI) (Sinclair & Moon 1997). This selection was based on the following criteria: (i) the dictionary is a phraseological or phraseology-related dictionary, (ii) it is a dictionary based on corpora, (iii) it is a reputable dictionary in terms of its publishing house, its editors and the lexicographers involved in its creation. However, before presenting the data and its related statistics, two questions need to be answered: What kinds of units are indexed in each dictionary? What lexicographic information is presented in the megastructure, macrostructure, and microstructure of each dictionary? In order to start answering those questions, the next section will offer a definition of megastructure, macrostructure, and microstructure.

2.1. Lexicographic information: megastructure, macrostructure, and microstructure

The present analysis is partly concerned with the ways in which PUs are represented in these two dictionaries. Thus, it is necessary to distinguish the characteristics of the sources from which data have been extracted. In order to do so, three parts of the dictionary had to be analyzed, namely: (i) the dictionary's megastructure, (ii) its macrostructure, and (iii) its microstructure. The definitions given by Hartmann & James (1998) for these three terms will be the ones adopted in this paper. According to these authors, the megastructure "includes the macrostructure and the outside matter" (1998: 93); the macrostructure is "the overall list structure which allows the compiler and the user to locate information" (1998: 91); finally, the microstructure is defined as "the internal design of a reference unit" (1998: 94).

The DFDEA's megastructure encompasses seven sections: (i) the motivation of the dictionary, (ii) the guidelines of use, (iii) a list of abbreviations used in the dictionary, (iv) a glossary of linguistic terms, (v) an alphabetical consultation guide, (vi) the body of the dictionary, and (vii) a list of cited texts. All this information comprises 1,084 pages.

The first section of the DFDEA, related to the motivation of this lexicographic work, explains the choosing of three words included in the title of the dictionary: *fraseológico* (phraseological), *documentado* (documented), and *actual* (current). According to its editors, the dictionary is *fraseológico* because it contains several types of PUs, including idioms, collocations, formulaic expressions, foreign-language idioms, and sayings (Seco, Andrés & Ramos 2004: xvi-xviii) as exemplified in Table 3.

Table 3. PU examples taken from the DFDEA

Type of PU	Example
Idiom	*callejón sin salida*
Collocation	*prestar atención*
Formulaic expression	*calladito estás mejor*
Idioms in other languages	*sine qua non*
Sayings	*a lo hecho, pecho*

In the DFDEA two types of sources were used in order to retrieve the phraseological entries: corpora and the press. The corpora used included two resources from the Real Academia Española (CORDE and CREA), one that was compiled for the *Diccionario del español actual* (Seco, Andrés & Ramos 1999), and one *ad hoc* corpus for this specific project. The authors do not add any further information about how newspapers were used for the extraction of PUs; however, the last part of the dictionary has an appendix that contains all the texts cited, including the press references that were used (Seco, Andrés & Ramos 2004: xiii-xiv).

Finally, its temporal aspect turns this lexicographic work into a synchronic dictionary. It was developed by using sources from a period spanning almost 50 years (1955 to 2004), thus offering a picture of phraseology up to that time.

It is worth mentioning, that there is one aspect that was not explained in depth in the first section of the DFDEA regarding how PUs were indexed in the dictionary. The authors only explain that PUs are listed under certain headwords. Those headwords are emphasized in the consultation guide through the use of bold letters (see Table 4).

Table 4. Examples of headwords in the DFDEA

Phraseological Unit	**Type of PU**	**Headword**
***hombre** de la calle*	Noun idiom	*hombre* (man)
*como un solo **hombre***	Adverbial idiom	
*hacer un **hombre***	Verbal idiom	
*vamos, **hombre***	Interjectional idiom	

Therefore, at first sight, it looks like the expressions are listed under the noun (when present.) However, after a further analysis of other examples (see Table 5) there is no evidence of any practical or theoretical motivation for choosing a word in particular.

Table 5. Incongruence of headword choosing in the DFDEA

Phraseological Unit	Type of PU	headword
*clamar al **cielo***	Verbal idiom	*cielo* (heaven)
*clamar en el **desierto***	Verbal idiom	*desierto* (desert)
***clamar** justicia*	Verbal idiom	*clamar* (to cry out)
***clamar** venganza*	Verbal idiom	

The CCDOI consists of four main sections: (i) the introduction, (ii) the guidelines of use, (iii) the body of the dictionary, and (iv) an alphabetic consultation index of the PUs. The dictionary length is 493 pages.

The first section of the CCDOI includes a detailed explanation of the motivation behind this lexicographic work, the sources used for the extraction of PUs, as well as the definition of *idiom*. Among the PUs that the authors extracted for their inclusion in the CCDOI one can find not only idioms but also a wide range of expressions (see Table 6.) However, it is stated that phrasal verbs such as "give up" or "put off" are not included in this work (Sinclair & Moon 1997: v).

Table 6. Examples of PUs included in the CCDOI

Type of PU	Example
Idiom	spill the beans
Multiword metaphors	the acid test
Metaphorical proverbs	in for a penny, in for a pound
Expressions with pragmatic meaning	famous last words

The main source for the extraction of PUs was the Bank of English, a subset of the Collins Corpus, containing approximately 650 million words (HarperCollins n.d.). One feature that this dictionary presents is the frequency-of-occurrence mark based on the corpora from where they were extracted. In the dictionary's introductory section, the editors explain that idioms have an infrequent level of occurrence in texts. The dictionary offers a scale in which the least frequent idioms occur less often than once per 10 million words and the most frequent at least once per two million words (Sinclair & Moon 1997: v). This scale is included in front of each idiom (see Table 7).

Table 7. Frequency bands in the CCDOI

PU	Frequency indicator	Range
prepare the ground	◀◀◀	Once every two million words
fire on all cylinders	◀◀	Not specified in the dictionary
all system go	◀	Between 1 and 3 times every 10 million words
come down in the world	No indicator	Not specified in the dictionary

The consultation index in the CCDOI is organized alphabetically, based on the headwords of the PUs. Another characteristic of this index is that it does not take the determiners 'the', 'a', or 'an' into consideration for the alphabetical-order distribution of the PUs (see Table 8).

Table 8. PU examples taken from the CCDOI

Headword	**PU order**
Light	a leading **light**
	light a fire under someone
	light as a feather
	the **light** at the end of the tunnel

At this point, it is noteworthy that the macrostructure of both the DFDEA and the CCDOI share the same lemmatization and indexation of entries as it can be seen in their own consultation guidelines. That means that PUs are listed under certain headwords, and, subsequently, those headwords are listed alphabetically. In contrast, the microstructures of both dictionaries differ in the information they include, as illustrated in Table 9.

Table 9. Lexicographic article in the DFDEA and the CCDOI (microstructure)

DFDEA	**CCDOI**
Headword	Headword
Entry (in alphabetical order)	Entry (in alphabetical order excluding determiners)
Grammatical/Functional marking	**Not included**
Diasistematical marking (colloquial, jargon, etc.)	**Not included**
Not included	Frequency band (according to the Bank of English)
Definition (direct or by context of use)	Definition (sentence like definition)
Example (Concordance from corpora or the press)	Example (Concordance from the Bank of English)
Source of the example	**Not included**

Since one of the objectives of this study is to analyze the morphosyntactic patterns related to idioms, the absence of markings (word class e.g.: nominal, adjectival, verbal, etc.) in the CCDOI was an obstacle for achieving such goal. Therefore, it was necessary to assign a grammatical/functional mark to each entry of the CCDOI. However, this procedure will be explained in detail in the next section of this paper.

As shown in Table 8, the two dictionaries of interest present definitions in different ways. In the DFDEA definitions are presented in a manner in which they can substitute the entry in certain contexts. In cases in which it is not possible to offer a "direct" definition, the dictionary includes an explanation of the use of the expression. The CCDOI include full-sentence definitions that explain the different contexts of use for each expression (see Table 10).

Table 10. Examples of definitions in the DFDEA and the CCDOI

Expression	Definition
pedir peras al olmo	*Esperar o pretender imposibles.* (Seco, Andrés & Ramos 2004: 774)
like getting blood out of a stone	If you have difficulty persuading someone to give you money or information, you can say that it is like getting blood out of a stone. (Sinclair & Moon 1997: 36)
Dios los cría y ellos se juntan	*Se usa para comentar la unión de personas de caracteres o intereses similares.* (Seco, Andrés & Ramos 2004: 398)
birds of a feather flock together	If you describe two or more people as birds of a feather, you mean that they are very similar in many ways. (Sinclair & Moon 1997: 34)

Determining if one way of defining the PUs is better than the other depends on each reader's —e.g. a linguist, translator, or enthusiast of phraseology— interests. What becomes apparent is that neither of those definitions could provide a quick solution for a user who does not know exactly which PU he/she is looking for. Nonetheless, a solution to this problem will be humbly proposed in the course of this paper.

2.2. Data selection and database compilation

In order to carry out the analyses proposed in this study, two databases were compiled (one containing the entries of the DFDEA, and the other containing the entries of the CCDOI). The database in Spanish includes 16,760 PUs composed by 55,831 word forms, while the database in English contains 4,285 PUs composed by 18,123 tokens. The tokens in both databases include grammar and lexical words as well as punctuation marks.

One limitation in phraseological studies aiming at characterizing sets of units has to do with the selection of the analytical sample. The amount of PUs included in the two databases for this study goes beyond what could be informed about in a single paper. Therefore, it was necessary to reduce the number of units for the analysis while maintaining a representative group of them. It was decided then to single out a limited number of selection criteria from the data starting with the number of forms (see Fig. 1).

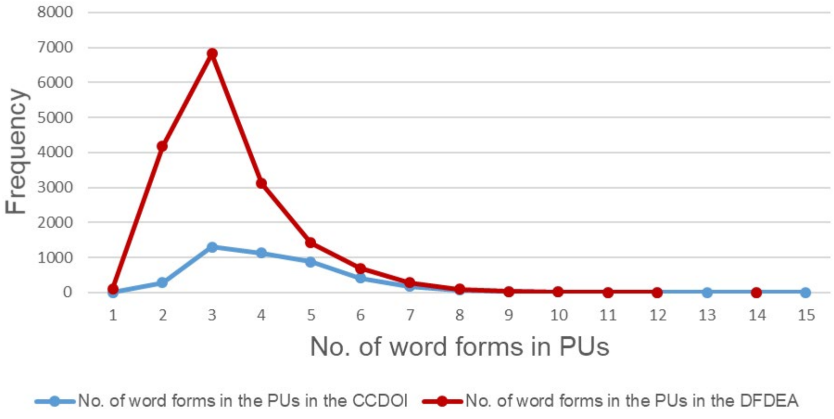


Figure 1. No. of word forms of the PUs in the DFDEA and the CCDOI.

It becomes apparent that Figure 1 shows a certain inconsistency with the criterion of plurilexicality of PUs. For example, a closer look to the data shows that the DFDEA indexed 105 PUs with only one word form, while the CCDOI indexed only one PU with one word form. The Spanish PUs were enclitic units, and, therefore, they were presented as one-form PUs (which is common use in lexicographic entries). Moreover, the only PU word form in English is an initialism: OTT (over the top) that was indexed in the CCDOI under the headword top (see Table 11).

Additionally, Figure 1 shows an uneven distribution of the frequency of PUs' number of forms. The only group that alters that frequency distribution is the one comprising phraseological bi-grams. One possible explanation for this might be related to an editorial decision to avoid these kinds of units due to the difficulty in drawing a boundary between compounds and PUs. This tendency has already been observed in other lexicographic resources in Spanish (Rojas Díaz & Pérez Sanchez 2019). However, reaching a satisfactory conclusion on the reasons behind this frequency distribution would only be possible through a deeper analysis of this group of units.

Table 11. Contexts of use of one-form PUs indexed in the DFDEA and in the CCDOI

Expression	Context of use
componérselas	*Tiene mucha familia y parece buen prójimo. Mal **se las** va a **componer** el hombre.* (Seco, Andrés, & Ramos 2004: 303).
OTT	Each design is very different in style. Some are subtle, some gloriously **OTT** (Sinclair & Moon 1997: 397).

With the information presented above, the criteria needed in order to carry out the sample selection was finally available. Thus, the first selection criterion was the number of forms. Then, the PUs consisting of three, four, and five forms were chosen. That selection, in turn, allowed for the study of more than 50% of the entries in both dictionaries (see Table 12).

Table 12. Distribution of PUs consisting of three - five forms, indexed in the DFDEA and in the CCDOI

No. of forms	Frequency in the DFDEA (%)	Frequency in the CCDOI (%)
Three-form PUs	6828 (40.7%) (*abogado del diablo*)	1294 (30.2%) (bite your tongue)
Four-form PUs	3117 (18.6%) (*cara de pocos amigos*)	1122 (26.2 %) (dig your own grave)
Five-form PUs	1422 (8.5%) (*el malo de la película*)	870 (20.3 %) (get your brain into gear)

Since this study has been conceived as the starting point of a larger project involving specialized dictionaries, the second criterion for sample selection was the functional marking of each PU. The DFDEA offered an extensive set of marks for this purpose; however, that was not the case for the CCDOI, as shown in Table 9, above. Given that this criterion was central both for this study, and, as said before, for further in-depth, specialized-lexicography studies, all the entries of the CCDOI were marked manually by using the marking set provided by the DFDEA (Seco, Andrés, & Ramos 2004: xxvii-xxviii) and the functional/grammatical information available in another dictionary related to the Cobuild project (Sinclair 2006). Once all the entries in both dictionaries were marked, 33 different marks were identified in the DFDEA, and 17 in the CCDOI.

The analysis of functional marking in both dictionaries allowed for the identification and selection of PUs' grammatical functions of interests. Thus, on the one hand, verb PUs were chosen for in-depth analysis because it was the most frequent mark in the DFDEA and in the CCDOI. On the other hand, given that authors such as Sager (1990: 58) and L'Homme (2004) assert that nouns are predominant in concept representation in specialized dictionaries, noun PUs were selected as the second category to be analyzed in depth. Once that selection was made, the analysis databases were finally set for carrying out the study on 4,932 PUs chosen from the DFDEA and 2,387 PUs from the CCDOI (See Fig. 2).

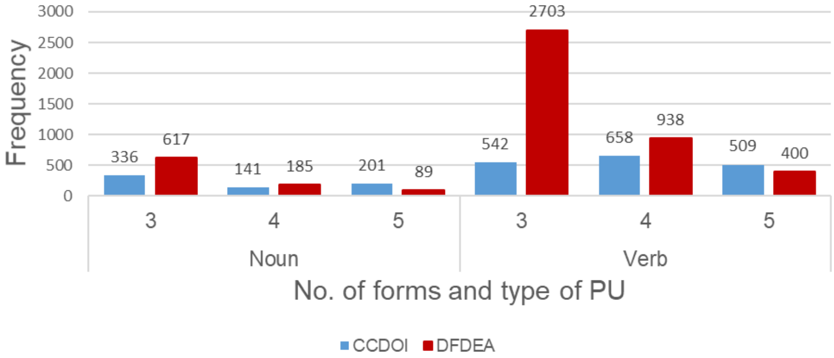


Figure 2. No. of nominal and verbal PUs in the analysis databases and their corresponding number of forms.

In summary, based on criteria such as frequency and importance for concept representation, verb and noun PUs were selected from both dictionaries in order to create the analysis databases for this study. Once the databases were set up, the units included in them were analyzed in depth as explained in the following section.

3. Analysis and Results

In view of the fact that the objective of this study is to extract as much linguistic information from the PUs as possible three different analyses were performed: (i) lexical, (ii) semantic, and (iii) morphosyntactic. Those analyses will be explained in detail here.

3.1. Lexical analysis

The first step in order to perform the lexical and the morphosyntactic analyses was to implement a Part-of-Speech (henceforth POS) tagging on the databases. For that task, the TreeTagger (Schmid 1994) was employed, followed by a homogenization of the tags in order for them to be readable. 26,277 forms were tagged including Saxon possessive morphemes (e.g. a baker's dozen) and hyphens (-) in English, constituted as categories (see Table 12).

Table 13. Distribution of component words by POS in the databases

POS	Frequency in the DFDEA (%)	Frequency in the CCDOI (%)
Noun	5,430 (32.14%)	3,354 (35.76%)
Verb	4,349 (25.74%)	1,731 (18.45%)
Determiner	2,876 (17.02%)	1,447 (15.43%)
Preposition	2,166 (12.82%)	1,233 (13.14%)
Adjective	667 (3.95%)	552 (5.88%)
Pronoun	198 (1.17%)	514 (5.48%)
Adverb	526 (3.11%)	139 (1.48%)
Contraction	318 (1.88%)	0 (0.00%)
Conjunction	213 (1.26%)	90 (0.96%)
Past Participle	132 (0.78%)	78 (0.83%)
Saxon possessive	0 (0.00%)	126 (1.34%)
Present Participle	7 (0.78%)	72 (0.77%)
Hyphen	0 (0.00%)	42 (0.45%)
Demonstrative	11 (0.07%)	0 (0.00%)
Interjection	4 (0.02%)	2 (0.02%)

The ‘noun’ category is the most frequent among the component words followed by the ‘verb’ category (see Table 13). This goes in contrast with the predominance of verbal PUs shown in Fig. 2. However, the reason for having more nouns than verbs in the word class counting is that a number of nouns co-occur with verbs in verb PUs.

Once POS frequency was determined, a word cloud was plotted in order to identify the most frequent nouns and verbs among the component words of the PUs (see Fig. 3).

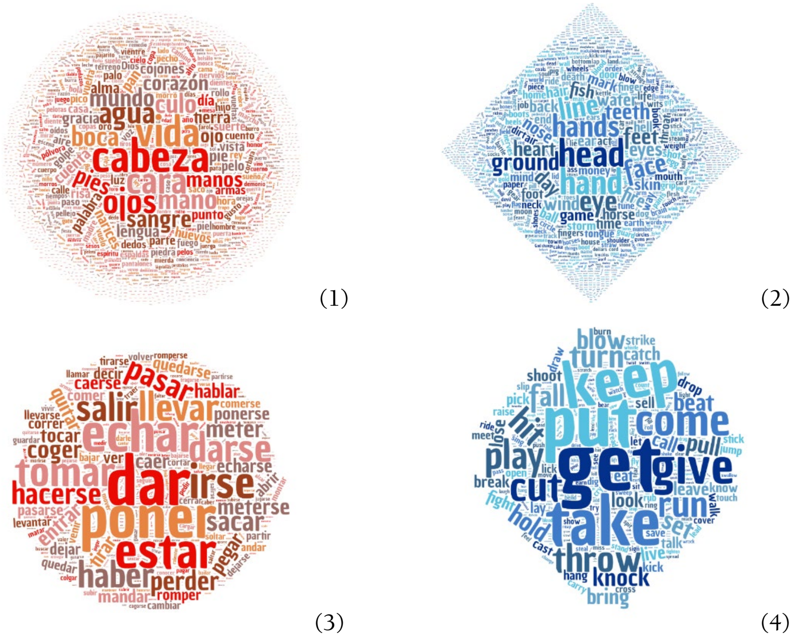


Figure 3. Nominal and verbal components of the PUs in the DFDEA (1 and 3) and in the CCDOI (2 and 4).

The word clouds presented in Figure 3 show the most frequent word (by form, not by lemma) in the center of each graph. The words' size in the graph is directly proportional to their frequency in the databases. The first interesting finding extracted from this representation of data is that the most frequent component words of the PUs in the databases are words used in everyday language. This was validated through a search in general language corpora, namely:

- Spanish: *Corpus diacrónico del español* (CORDE) (Real Academia Española n.d.) and *Corpus del español* (*genre/historical*) (Davies 2002).
- English: word frequency lists based on the British National Corpus (Leech, Rayson, & Wilson 2001) (Kilgarriff 2006)].

All the words in the database were lemmatized. Thus, it becomes apparent that although the most frequent word form in the word cloud (see Fig. 3) is *cabeza,* once the database was lemmatized, this information could vary (see Table 14). The table below illustrates two different scores. On the one hand, the CORDE (Real Academia Española, n.d.), and the frequency lists by Kilgarriff (2006) show the ranks of such words in a corpus. On the other hand, the *Corpus del español* (2002) and the frequency lists by Leech, Rayson & Wilson (2001) provide ranks based on each word's POS. It is evident that all the words in the databases are ranked within the top 1000 most frequent words in Spanish and in English respectively, according to the corpora consulted. The words in Table 13 are also within the top 200 most frequent nouns and verbs according to the POS frequency list of the *Corpus del español* (2002) and the frequency lists by Leech, Rayson & Wilson (2001).

Table 14. Top 5 ranking for nouns and verbs of component lemmas of PUs in corpora

Spanish (DFDEA)				**English (CCDOI)**			
POS	**Lemma**	**CORDE**	***Corpus del español***	**POS**	**Lemma**	**Kilgarriff**	**Leech, Rayson, & Wilson**
Noun	*ojo*	184	14	Noun	hand	176	26
	mano	158	26		eye	240	43
	cabeza	324	47		head	241	38
	vida	99	5		foot	484	163
	cara	522	112		line	278	73
Verb	*dar*	136	40	Verb	get	44	8
	hacer	140	17		have	8	2
	tener	192	36		go	40	10
	ser	51	6		put	125	26
	poner	387	119		take	54	13

Table 14 presents the top 5 nouns and verbs extracted from the databases. This gives indications about the POS and semantic-category distribution of the words in the databases.

3.2. Semantic analysis

The second type of analysis performed in this study is a semantic one. In recent years, scholars have progressively explored some semantic aspects of phraseology, more especially in studies related to terminology and languages for specific purposes (Grčić Simeunović & de Santiago 2016; Patiño 2017). For this study, however, a different semantic approach was taken. The UCREL's Semantic Analysis System (henceforth USAS) was employed. USAS is a POS and semantic tagger, containing semantic tags divided into 232 semantic categories based, in turn, on 21 discourse fields identified by McArthur (1981) (Archer, Wilson & Rayson 2002: 2).

All the word forms of the database were tagged with this semantic tagset and revised and corrected manually in both languages, thus creating four analysis layers, namely: lexical, grammatical, discourse filed, and semantic category. These four layers made it possible to observe how certain morphosyntactic patterns interacted with different sequences of semantic categories (hereinafter semantic patterns) (see Table 15).

Table 15. Database tagging sample in Spanish and English

	Spanish (DFDEA)			
Word form	*el*	*ombligo*	*del*	*mundo*
POS	Prep	N	Contr	N
Discourse field	Z	B	Z	W
(Descriptive)	Names and grammatical words	Body and the individual	Names and grammatical words	World and the environment
Semantic level	Z5	B1	Z5	W1
(Descriptive)	Grammatical bin	Anatomy and physiology	Grammatical bin	The universe
	English (CCDOI)			
Word form	throw	in	the	towel
POS	V	Prep	Det	N
Discourse field	M	Z	Z	B
(Descriptive)	Movement, location, travel, and transportation	Names and grammatical words	Names and grammatical words	Body and the individual
Semantic level	M2	Z5	Z5	B5
(Descriptive)	Putting, pulling, pushing, transporting	Grammatical bin	Grammatical bin	Clothes and personal belongings

Some information —such as the distribution of word forms in semantic fields— could only be obtained when the databases were tagged by using the USAS tagset. When comparing the information from Fig. 3 with that presented in Table 13, it is possible to state that there is a strong tendency for parts of the body to occur as a word form in the database. Nevertheless, when comparing the whole distribution of word forms in the databases, according to the discourse fields provided by McArthur (1981) and tagged through USAS, it is possible to observe that the category "the body and the individual" ranks fourth (see Table 16). As observed elsewhere, in a lexicographic resource in Spanish, the occurrence of these words used in everyday language is an indicator of embodiment in the creation and fixation of PUs in general language

(Rojas Díaz & Pérez Sanchez 2019: 9). Embodiment is a concept developed in cognitive linguistics, and it is based on the statement that "our concepts, our ideas are influenced and composed by the structure of our bodies, by our own experience of the world that surrounds us" (Ibarretxe-Antuñano & Valenzuela 2016: 44, author's translation)

Table 16. Distribution of discourse fields tags in the databases

Discourse field	Frequency in DFDEA (%)	Frequency in CCDOI (%)
Names and grammar (Z)	6,249 (36.98%)	3,958 (42.39%)
General and abstract terms (A)	2,343 (13.87%)	945 (10.12%)
Movement, location, travel, and transportation (M)	1,979 (11.71%)	800 (8.57%)
Body and the individual (B)	1,431 (8.47%)	665 (7.12%)
Substances, materials, objects, and equipment (O)	763 (4.52%)	657 (7.04%)
Social actions, states, and processes (S)	641 (3.79%)	204 (2.18%)
Numbers and measurement (N)	527 (3.12%)	343 (3.67%)
Psychological actions, states, and processes (X)	466 (2.76%)	239 (2.56%)
Life and living things (L)	455 (2.69%)	300 (3.21%)
Language and communication (Q)	356 (2.11%)	169 (1.81%)
Food and farming (F)	320 (1.89%)	158 (1.69%)
Emotion (E)	294 (1.74%)	120 (1.29%)
World and environment (W)	221 (1.31%)	109 (1.17%)
Government and public (G)	182 (1.08%)	97 (1.04%)
Entertainment, sports, and games (K)	169 (1.00%)	146 (1.56%)
Architecture, housing and home (H)	165 (0.98%)	112 (1.20%)
Time (T)	161 (0.95%)	98 (1.05%)
Money and commerce in industry (I)	105 (0.62%)	184 (1.97%)
Science and technology (Y)	30 (0.18%)	5 (0.05%)
Arts and crafts (C)	27 (0.16%)	19 (0.20%)
Education (P)	13 (0.08%)	10 (0.11%)

Although this study is not intended to do a contrastive analysis between languages, but to present the information in parallel, some contrastive insights could be obtained when taking a closer look at the databases. There is a tendency in both dictionaries for certain types of words to occur within specific discourse fields, as is the case for verbs indicating movement (7.85% in the databases), nouns related to body parts (7.04% in the databases), and adjectives describing measurements (1.42% in the databases) (see Fig. 4).

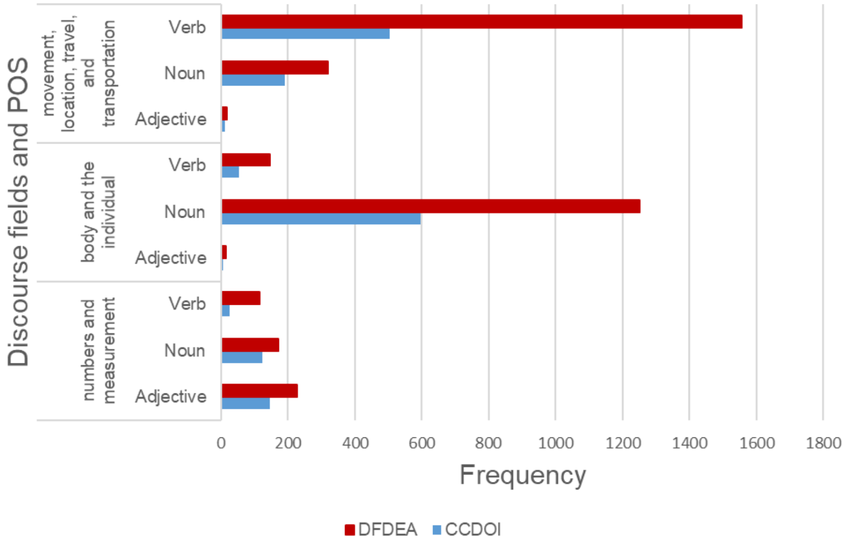


Figure 4. Distribution of main lexical categories (verbs, nouns, and adjectives) grouped by discourse fields.

At first sight, most semantic relationships between PUs' word forms and their meanings seem to be metaphorical. Nevertheless, a closer look shows that several cases are also metonymical (see Table 17), but an in-depth analysis of such semantic relationships is beyond the scope of the present work.

Table 17. Examples of metaphorical and metonymical relationships between PUs and their meaning

DFDEA		CCDOI	
Phraseological Unit	**Meaning**	**Phraseological Unit**	**Meaning**
arma de doble filo (Metaphorical)	*Cosa, argumento o procedimiento, que puede ocasionar un resultado opuesto al que se pretende.*	the salt of the earth (Metaphorical)	If you describe someone as the salt of the earth, you are showing admiration for their honesty
cargar la barriga (Metonymical)	*Quedarse embarazada*	give someone a black eye (Metonymical)	If you give someone a black eye, you punish them severely for something they have done, but without causing them permanent harm

Finally, it is necessary to state that semantic analysis is crucial not only for the type of work intended here, but also for phraseological studies in general, given that the study of meaning in phraseology will shed light on both the understanding of PUs and on their proper representation in lexicographical resources.

3.2. Morphosyntactic analysis

The third analysis in this study was morphosyntactic. 816 morphosyntactic patterns were extracted from the DFDEA and CCDOI. 388 (47.5%) of those patterns had two or more occurrences in the database. When combining both variables, it was possible to make a query about morphosyntactic and semantic patterns among the PUs. Table 18 includes the top-five most frequent patterns for noun and verb PUs in the dictionaries.

Table 18. Top five morphosyntactic patterns of noun and verb PUs in the DFDEA and the CCDOI

Morphosyntactic pattern	Frequency	Example	Type of PU	Dictionary
V Det N	1567	*abandonar el barco*	Verb	*DFDEA*
V Prep N	422	*bailar de alegría*		
V Prep Det N	398	*caber en la cabeza*		
V Contr N	150	*dar del vientre*		
V Adj N	94	*echar buen pelo*		
N Prep N	393	*faena de aliño*	Noun	
N Prep Det N	63	*gatos en la barriga*		
N Contr N	59	*hombre del saco*		
Det Adj N	37	*la mínima expresión*		
Det N Prep Det N	37	*un cero a la izquierda*		
V Det N	233	jump the gun	Verb	CCDOI
V Pron N	134	keep your cool		
V Prep Det N	102	lay down the law		
V Det N Prep N	90	make a meal of something		
V N Prep Det N	69	play things by the book		
Det N N	109	a bean counter	Noun	
Det Adj N	108	the acid test		
Det N Prep Det N	71	a skeleton in the cupboard		
Det N Prep N	48	the kiss of death		
Det PrP N	26	a sitting duck		

It is possible to nest the morphosyntactic patterns with the semantic patterns, which makes it possible to identify nouns related to certain discourse fields. In Table 19, some examples of the semantic patterns linked to the most frequent morphosyntactic patterns are shown. Each of the letters composing the semantic patterns corresponds to one of the discourse field labels previously presented in Table 16. From another point of view, it is also possible to look for semantic patterns and to observe morphosyntactic patterns that follow a specific semantic combination.

Table 19. Example of pattern nesting of semantic tags in morphosyntactic patterns

Morphosyntactic pattern	**Semantic pattern**	**Example**	**Type of PU**	**Dictionary**
V Det N	M Z B	*alzar la hombro*	Verb	*DFDEA*
	B Z B	*cagarse los calzones*		
N Prep N	B Z O	*lengua de trapo*	Noun	
	Q Z S	*cuento de hadas*		
V Det N	E Z H	hit the roof	Verb	*CCDOI*
	X H O	know the ropes		
Det Adj N	Z O O	a bright spark	Noun	
	Z O Q	a dirty word		

Similarly, it is also possible to nest the PUs by taking semantic patterns as a starting point, and then looking at what morphosyntactic patterns can be derived from those semantic patterns. This means that semantic tags could be used as a variable for either the extraction or classification of PUs in corpora, or for the indexation in lexicographic resources (see Table 20).

Table 20. Example of pattern nesting of morphosyntactic tags in semantic patterns

Semantic pattern	Morphosyntactic pattern	Example	Type of PU	Dictionary
M Z B	V Det N	*correr la sangre*	Verb	*DFDEA*
	V Prep N	*caerse de culo*		
	V Contr N	*salir del corazón*		
	V Prep V	*echarse a dormir*		
	V Prep PP	*ir de dormida*		
	V Prep Adj	*pasar a limpio*		
B Z O	N Prep N	*lengua de trapo*	Noun	
	N Contr N	*cana al aire*		
	N Adv N	*ojos como platos*		
	N Conj N	*pelos y señales*		
A Z O	V Det N	*fan the flames*	Verb	*CCDOI*
	V Prep N	*be in overdrive*		
	V Pron N	*blow your stack*		
	V N N	*give someone stick*		
	V Conj V	*crash and burn*		
	V Adj N	*spread like wildfire*		
	V N Adj	*catch someone cold*		
Z O O	Det Adj N	*a black mark*	Noun	
	Der N N	*the brass ring*		

The count of morphosyntactic and semantic patterns in the databases is presented in the following table. The information includes the frequency and relative percentages of each of the patterns that were extracted (see Table 21).

Table 21. Summary of the frequencies and percentages of the patterns in the databases

Dictionary	Type of PU	Morphosyntactic pattern		Semantic patterns	
DFDEA	Noun	124	(29.5 %)	461	(33.8 %)
	Verb	296	(70.5 %)	904	(66.2 %)
	Total DFDEA	**420**		**1,365**	
CCDOI	Noun	126	(31.8 %)	455	(35.3 %)
	Verb	270	(68.2 %)	835	(64.7 %)
	Total CCDOI	**396**		**1,290**	
Total dictionaries		***816***	***(23.5%)***	***2,655***	***(76.5 %)***
Grand total		*3,471*			

Traditionally, the morphosyntactic patterns resulting from the analyses (that have been carried out in the phraseology studies) have been used for the recognition and the extraction of candidates of PU in corpora. Nevertheless, the results regarding the nesting of morphosyntactic patterns and semantic patterns offered in this article will shed light on how to enhance the recognition method through semantic annotation but also the analysis of metaphorical and metonymical patterns.

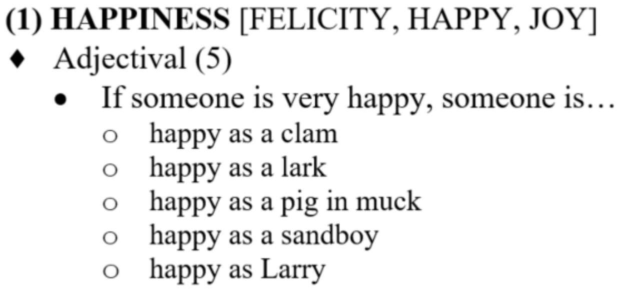
4. Discussion, Conclusions, and a Practical Proposal

This study offers 3,471 patterns. They are divided as follows: 816 (23.5%) are morphosyntactic patterns (420 from the DFDEA and 396 from the CCDOI) and 2,655 (76.5%) are semantic patterns (1,365 from the DFDEA and 1,290 from the CCDOI). The distribution of these patterns follow a ratio of almost 1:3, meaning that for each morphosyntactic pattern extracted three semantic patterns were identified. This ratio (1:3) is consistent within the dictionaries and the different types of idioms. The most frequent morphosyntactic patterns and semantic patterns of nominal and verbal units consisting of three, four, and five forms were exemplified and presented. This information can be used as a gold standard in order to make a comparison between the linguistic features found in PUs in general language dictionaries and PUs in specialized dictionaries.

Semantic tagging of databases or corpora offers the opportunity for testing different parameters for the extraction of PUs aiming at lexicographic or terminographic work. The use of morphosyntactic and semantic annotation for PUs opens the discussion on how PUs should be indexed in dictionaries nowadays. Although semasiology and onomasiology are two very well-known concepts in lexicography and terminography, it is evident that most phraseological dictionaries follow a semasiological approach for the indexation of entries, i.e., those dictionaries answer the question of what does X (word/phrase/idiom/proverb) mean? However, such approach requires the user to know the form or the expression he/she is looking for (Kocjančič 2004). Additionally, the results of an analysis like the one presented here can also provide empirical data useful for the study of the semantic composition of metaphorical and metonymical constructions.

The frequency analysis along with the semantic information extracted from the component words of the PUs of the DFDEA and the CCDOI shows the use of common words of our daily experiences to describe more complex conceptions through rhetorical devices such as similes, metonymies and metaphors as it has already been observed in several studies in corpora and lexicographic resources. (Ellis 2008; Sharma 2018; Torijano & Recio 2019; Rojas Díaz & Pérez Sanchez 2019). The results of the previously mentioned analyses support some of the views of Cognitive Semantics regarding the embodiment hypothesis (Ibarretxe-Antuñano & Valenzuela 2016: 37).

In many cases, dictionary users (e.g. translators) do not know the exact form or expression they are looking for. As a result, looking for similar or equivalent expressions in semasiological dictionaries becomes a time-consuming task. One solution that could be offered to users so that they can do better and more efficient searches for PUs in dictionaries would be to transform the way in which dictionaries present phraseological entries. That could be done by grouping PUs' entries semantically, following a hybrid indexation that uses semasiological and onomasiological approaches. Therefore, suggestion derived from the present study entails the display of information in phraseological entries somehow as it is exemplified in Figure 5

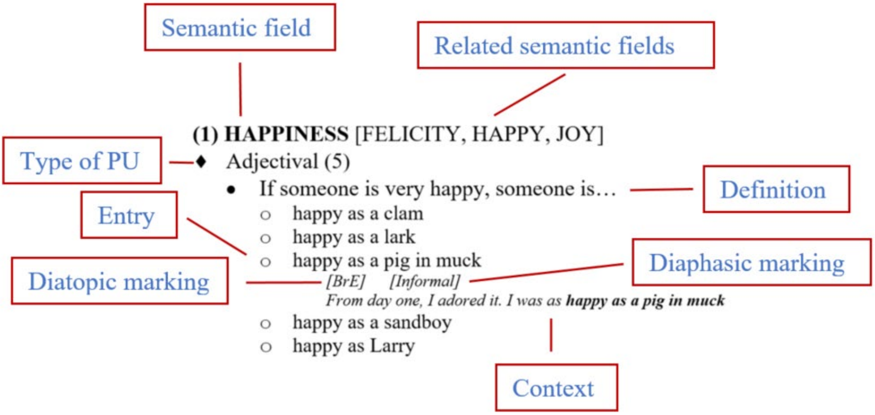


(1) HAPPINESS [FELICITY, HAPPY, JOY]

♦ Adjectival (5)
- If someone is very happy, someone is…
  - happy as a clam
  - happy as a lark
  - happy as a pig in muck
  - happy as a sandboy
  - happy as Larry

Figure 5. Example of an entry with an onomasiological/semasiological hybrid approach related to happiness.

The five idioms presented in Figure 4 have the same meaning in the CCDOI "If you are X, you are very happy". A representation like the one in Figure 5 not only allows the user to look for the expression needed, but it also provides the user with similar expressions. The lexicographic article could be expanded in order to provide the user with more lexicographic information such as diatopic marking (related to the place), diaphasic marking (related to language register), and contexts, among others (see Fig. 6).



(1) HAPPINESS [FELICITY, HAPPY, JOY]

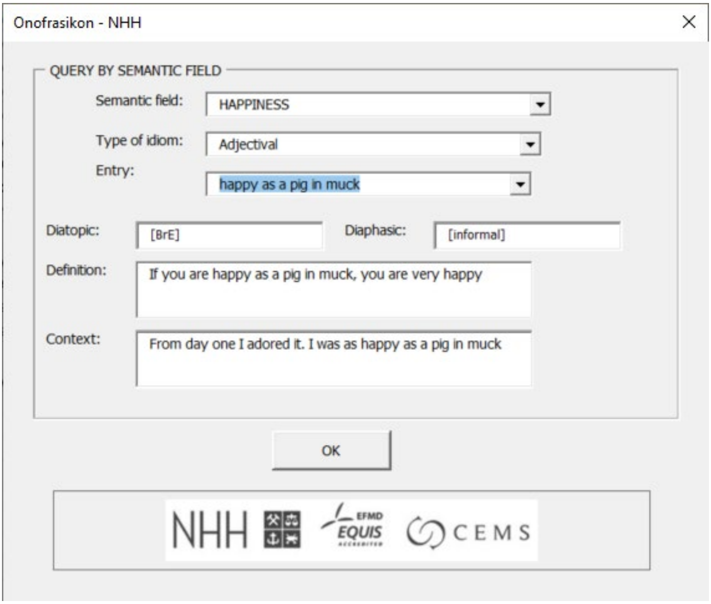
♦ Adjectival (5)
- If someone is very happy, someone is…
  - happy as a clam
  - happy as a lark
  - happy as a pig in muck

    [BrE] *[Informal]*

    From day one, I adored it. I was as ***happy as a pig in muck***
  - happy as a sandboy
  - happy as Larry

Figure 6. Example of an entry with an onomasiological/semasiological hybrid approach.

Information could also be presented in an electronic format (see Fig. 7) allowing the user to make different kind of queries if the entries are annotated morphosyntactically and syntactically.



Onofrasikon - NHH

QUERY BY SEMANTIC FIELD

Semantic field: HAPPINESS

Type of idiom: Adjectival

Entry: happy as a pig in muck

Diatopic: [BrE]

Diaphasic: [informal]

Definition: If you are happy as a pig in muck, you are very happy

Context: From day one I adored it. I was as happy as a pig in muck

Figure 7. Example of an entry with an onomasiological/semasiological hybrid approach in an electronic format.

Evidently, this reflection on the lexicographic techniques used to compile dictionaries needs to be broadened and verified with users and lexicographers to test its suitability as a possible approach for the enhancement of the compilation of dictionaries.

Finally, although it is true that it is impossible to offer the whole picture of the paradigm of phraseology for a language on the basis of the analysis of dictionaries, the information, statistics, and findings presented here can be used as a starting point for a transformation in the description and indexation of PUs in future studies and projects related to phraseology and lexicography.

References

Aguado de Cea, Guadalupe. (2007) "A multiperspective approach to specialized phraseology: Internet as a reference corpus for phraseology." In: Posteguillo, Santiago; María José Esteve & María Lluïsa Gea Valor (eds.) 2007. *The Texture of Internet: Netlinguistics in Progress*. Newcastle: Cambridge Scholars Publishing. 182-207.

Alonso, Margarita (ed.) (2006) *Diccionarios y fraseología.* La Coruña: Universidade da Coruña.

Archer, Dawn; Adrew Wilson & Paul Rayson. (2002) *Introduction to the USAS category system.* Benedict project report.

Atkins, Sue & Michael Rundell. (2008) *The Oxford Guide to Practical Lexicography.* New York: Oxford University Press.

Bloomfield, Leonard. (1933) *Language.* New York: Henry Holt.

Buendía Castro, Miriam & Pamela Faber. (2015) "Phraseological units in English-Spanish legal dictionaries: a comparative study." *Fachsprache* 3:4, pp. 161-175.

Bushnaq, Tatiana. (2015) "A Retrospective Analysis of the Term Phraseological Unit." In: Boldea, Iulian (ed.) 2015. *Debates on Globalization. Approaching National Identity through Intercultural Dialogue*. Tîrgu-Mureş: Arhipelag XXI Press, pp. 167-176.

Carneado, Zoila & Antonia Trista. (1985) *Estudios de Fraseología.* La Habana: Academia de Ciencias de Cuba, Instituto de Literatura y Lingüística.

Casares, Julio. (1950) *Introducción a la lexicografía moderna*. Madrid: CSIC.

Castillo Carballo, María Auxiliadora. (2006) *El lema: tipos de entradas*. Electronic version: <https://www.liceus.com/producto/lema-tipos-entradas/>

Chafe, Wallace. (1968) "Idiomaticity as an Anomaly in the Chomskyan Paradigm". *Foundations of Language 4*, pp. 109-127.

Chomsky, Noam. (1965) *Aspects of the Theory of Syntax.* Cambridge: MIT Press.

Corpas Pastor, Gloria. (1996) *Manual de fraseología española.* Madrid: Gredos.

Cowie, Anthony. (2001) "Introduction" In: Cowie, Anthony (ed.) 2001. *Phraseology: Theory, Analysis, and Application.* Oxford: Oxford University Press, pp. 1-22.

Davies, Mark. (2002) *Corpus del español: 100 million word, 1200s-1900s*. (Corpus) Electronic version: <http:///corpusdelespanol.org/hist-gen>

Ellis, Nick. (2008) "Phraseology: The periphery and the heart of language." In: Meunier, Fanny & Sylviane Granger (eds.) 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, pp. 1-13.

Firth, John. (1957) *Papers in Linguistics, 1934-1951*. London: Oxford University Press.

García-Page, Mario. (2008) *Introducción a la fraseología española.* Barcelona: Anthropos.

González, María Isabel. (2006) "La definición lexicográfica de las unidades fraseológicas: la aplicación de modelos formales." In: Alonso, Maragarita (ed.) 2006. *Diccionarios y fraseología.* La Coruña: Universidade da Coruña, pp. 221-233.

Grčić Simeunović, Larisa & Paula de Santiago. (2016) "Semantic approach to Phraseological Patterns in Karstology." In: Margalitadze, Tinatin & George Meladze (eds.) 2008. *Proceedings of the XVII Euralex International Congress*. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 685-693.

HarperCollins. (n.d.) *The Collins Corpus*. (Corpus) Electronic version: <https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus>

Hartmann, Reinhard & Gregory James. (1998) *Dictionary of Lexciography.* New York: Routledge.

Heid, Ulrich. (2008) "Computational phraseology: An overview." In: Granger, Sylviane & Fanny Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins Publishing Company, pp. 337–360.

Hockett, Charles. (1958) *A Course in Modern Linguistics*. New York: Macmillan.

Householder, Fred. (1959) "On Linguistic Primes." *Word* 15, pp. 231-239.

Ibarretxe-Antuñano, Irialde & Javier Valenzuela (eds.) (2016) *Lingüística Cognitiva*. Barcelona: Anthropos.

idiom. (n.d.) *Cambridge dictionary*. Electronic version: <https://dictionary.cambridge.org/dictionary/english/idiom>

Jackendoff, Ray. (1997) *The Architecture of Language Faculty.* Cambridge: The MIT Press.

Kilgarriff, Adam. (2006) *BNC database and word frequency lists*. (Word list) Electronic version: <http://www.kilgarriff.co.uk/bnc-readme.html>

Kocjančič, Polonca. (2004) "Acerca de la macroestructura y la microestructura en el diccionario bilingüe." *Verba Hispanica* 12:1, pp. 171-185.

Leech, Geoffrey; Paul Rayson & Andrew Wilson. (2001) *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Leroyer, Patrick. (2006) "Dealing with phraseology in business dictionaries: focus on dictionary functions - not phrases." *Linguistik online* 27:2, pp. 183-194 Electronic version: <https://bop.unibe.ch/linguistik-online/article/view/750/1279>

L'Homme, Marie-Claude. (2004) *La terminologie: principes et techniques*. Montreal: Presses de l'Université de Montréal.

locución. (n.d.) *Dicionario de la lengua española*. Electronic version: <https://dle.rae.es/?id=NYSj8PH>

López, Xavier Pascual. (2012) *Fraseología española de origen latino y motivo grecorromano*. Lleida: Universitat de Lleida. PhD Thesis.

McArthur, Tom. (1981) *Longman Lexicon of Contemporary English*. London: Longman.

Mel'čuk, Igor. (2012) "Phraseology in the language, in the dictionary, and in the computer." In: Kuiper, Koenraad (ed.) 2012. *Yearbook of Phraseology* Vol. 3. New York: De Gruyter Mouton, pp. 31-56.

Mellado Blanco, Carmen. (2008) *Colocaciones y fraseología en los diccionarios*. Frankfurt am Main: Peter Lang.

Molina Plaza, Silvia. (2005) "English and Spanish phraseology in contrast." *RAEL: revista electrónica de lingüística aplicada*, pp. 174-189.

Moon, Rosamund. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. New York: Oxford University Press.

Moon, Rosamund. (2008) "Dictionaries and collocation." In: Granger, Sylviane & Fanny Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins, pp. 313-336.

Moon, Rosamund. (2009) *Words, Grammar, Text: Revisiting the Work of John Sinclair*. Amsterdam: John Benjamins.

Norrick, Neal. (2007) "English Phraseology". In: Burger, Harald; Dmitrij Dobrovol'skij & Peter Kühn (eds.) 2007. *Phraseology* Vol. II. New York: Walter de Gruyter, pp. 615-619.

Ortega Ojeda, Gonzalo & María Isabel González Aguilar. (2008) "La técnica fraseográfica: el DRAE-2001 frente al DEA-1999." In: Mellado, Carmen (ed.) 2008. *Colocaciones y fraseología en los diccionarios*. Frankfurt am Main: Peter Lang. pp. 232-245.

PAQUOT, Magali. (2015) "Lexicography and Phraseology." In: Biber, Douglas & Randi Reppen (eds.) 2015. *The Cambridge Handbook of Corpus Linguistics.* Electronic version: <https://dial.uclouvain.be/pr/boreal/en/object/boreal%3A139795/datastream/PDF_01/view>

PATIÑO, Pedro. (2017) *Description and representation in language resources of Spanish and English specialized collocations from Free Trade Agreements.* Bergen: NHH Norwegian School of Economics. PhD Thesis.

PENADÉS, Inmaculada. (2006) "La información gramatical sobre la clasificación de las locuciones en los diccionarios." In: Alonso, Maragita (ed.) 2006. *Diccionarios y fraseología.* La Coruña: Universidade da Coruña. pp. 249-259.

RAKOTOJOELIMARIA, Agathe. (2004) *Esbozo de un diccionario de locuciones verbales español-malgache.* Alcalá de Henares: Universidad de Alcalá. PhD Thesis.

REAL ACADEMIA ESPAÑOLA. (n.d.) *Banco de datos CORDE.* (Corpus) Electronic version from Corpus diacrónico del español: <http://www.rae.es>

ROJAS DÍAZ, José Luis & Juan Manuel Pérez Sánchez. (2019) "You Took the Word Out of My Mouth: a Morphosyntactic and Semantic Analysis of a Phraseological Lexicon of Colombian Spanish." In: Corpas Pastor, Gloria & Ruslav Mitkov (eds.) 2019. *Computational and Corpus-Based Phraseology.* Berlin: Springer, pp. 375-390.

RUIZ GURILLO, Leonor. (1997) *Aspectos de la fraseología teórica española.* Valencia: Universidad de Valencia.

RUIZ GURILLO, Leonor. (1998) *La fraseología del español coloquial.* Barcelona: Ariel.

RUIZ GURILLO, Leonor. (2001) *Las locuciones en el español actual.* Madrid: Arco Libros.

SAGER, Juan Carlos. (1990) *A Practical Course in Terminology Processing.* Amsterdam: John Benjamins.

SANTAMARÍA PÉREZ, Isabel. (2003) "Localización de la información fraseológica en el diccionario: cuándo y cómo situarla en la macroestructura." In: Alemany Bay, Carmen; Beatriz Aracil Varón; Remedios Mataix Azuar; Pedro Mendiola Oñate; Eva Valero Juan & Abel Villaverde Pérez (eds.) 2003. *Con Alonso Zamora Vicente: Actas del Congreso Internacional "La Lengua, la Academia, lo Popular, los Clásicos, los Contemporáneos...".* Alicante: Universidad de Alicante. pp. 1045-1057.

SCHMID, Helmut. (1994) "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester.

SECO, Manuel; Andrés Olimpia & Gabino Ramos (eds.) (1999) *Diccionario del español actual*. Madrid: Aguilar.

SECO, Manuel; Andrés Olimpia & Gabino Ramos (eds.) (2004) *Diccionario fraseológico documentado del español actual*. Madrid: Aguilar.

SHARMA, Sunil. (2018) "Happiness and metaphors: a perspective from Hindi phraseology." *Yearbook of Phraseology* 8:1, pp. 171-190.

SINCLAIR, John. (1984) "Lexicography as an Academic Subject." In: Hartmann, Reinhard (ed.) 1984. *LEXeter '83 Proceedings*. Tübingen: Niemeyer. pp. 3-12.

SINCLAIR, John. (2006) *Collins Cobuild Advanced Learner's English Dictionary*. Glasgow: HarperCollins.

SINCLAIR, John & Rosamund Moon (eds.) (1989) *Collins COBUILD Dictionary of Phrasal Verbs*. London: Harper Collins.

SINCLAIR, John & Rosamund Moon (eds.) (1997) *Collins COBUILD Dictionary of Idioms*. Glasgow: HarperCollins publishers.

SKOLNÍKOVÁ, Pavlína. (2010) *Las colocaciones léxicas en el español actual*. Brno: Masarykova univerzita. PhD Thesis.

SOSIŃSKI, Marcin. (2006) *Fraseología comparada del polaco y del español: su tratamiento en los diccionarios bilingües*. Granada: Universidad de Granada. PhD Thesis.

TORIJANO, José Agustín, & María Ángeles Recio. (2019) "Translating Emotional Phraseology: A Case Study." In: Corpas Pastor, Gloria & Ruslav Mitkov (eds.) 2019. *Computational and Corpus-Based Phraseology*. Berlin: Springer, pp. 391-403.

TSCHICHOLD, Cornelia. (2008) "A computational lexicography approach to phraseologisms." In: Granger, Sylviane & Fanny Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Bejamins. pp. 361–376.

ZULUAGA, A. (1980) *Introducción al estudio de las expresiones fijas*. Frankfurt: Peter Lang.

BIONOTE / BIONOTA

José Luis Rojas Díaz is a PhD Scholar who joined NHH Norwegian School of Economics (Norway) in 2017. He holds a Master in Linguistics from the University of Antioquia (Colombia) (2014). His research focuses on the representation of phraseological units in Spanish and English dictionaries. His study involves dictionaries from general language and specialized fields. His major research interests cover phraseology, lexicography, linguistics, and translation.

José Luis Rojas Díaz es un becario doctoral que en 2017 se unió a la *NHH Norwegian School of Economics* (Noruega). Es Magíster en Lingüística de la Universidad de Antioquia (Colombia) (2014). Su investigación se enfoca en la representación de unidades fraseológicas en diccionarios en inglés y español. Sus estudios incluyen diccionarios de lengua general y lenguajes especializados. Sus áreas de interés en investigación abarcan temas relacionados con la fraseología, la lexicografía, la lingüística y la traducción.