Spanish multiword expressions: looking for a taxonomy

Carla Parra Escartín
Almudena Nevado LlopisEoghan Sánchez Martínez

Thursday 22nd February, 2018

1 Introduction

Research on multiword expressions (s) has a long history both in linguistics and in Natural Language Processing (). Many researchers have addressed the challenge from different perspectives (**Melcuk:1987**; **Church:1990**; **Sinclair:1991**; **smadja1993**; **moon1998**; **Lin:1999**).

check tabulars in tabulars

s are part of the lexicon of native speakers of a language and thus are interesting from a theoretical linguistics point of view. Researchers working on language acquisition also assess the acquisition of s (**Devereux:2007**; **Villavicencio:2012**; **Nematzadeh:2013**); and they have also been researched in psycholinguistics (**Rapp:2008**; **Holsinger:2013a**; **Holsinger:2013b**; **Schulteimwalde:2015**), among other theoretical fields. In the case of applications, s need to be correctly detected and processed. In addition, when applications deal with two or more languages, the treatment of s needs to deal with multilingual aspects.

A lot of research has focused on specific subclasses of s (e.g. *idioms*, *collocations*, *light verb constructions*). More general works studying the phenomenon as such have focused on English, or have taken prior research on English as a starting point. However, this English-driven analysis needs to be further investigated taking other languages into account. As the intrinsic characteristics of a language vary, it seems necessary to use broad, general taxonomies that allow for the classification, description and analysis of s notwithstanding the language they are applied to. In this article, we test this by analyzing Spanish s using an existing taxonomy.

As a starting point of our study, we take the taxonomy proposed by **Ramisch:2012**; **Ramisch:2015**. He distinguishes three morphosyntactic classes and three additional so-called "difficulty classes". The three morphosyntactic classes are *nominal expressions*, *verbal expressions* and *adverbial and adjectival expressions*. Nominal expressions are further subdivided in *noun compounds*, *proper names* and *multiword terms*, and verbal expressions in *phrasal verbs* and *light verb constructions*. Finally, he distinguishes three difficulty classes: *fixed expressions*, *idiomatic expressions*, and *"true" collocations*.

We created a data set of Spanish s with the aim of finding examples of each type of proposed by **Ramisch:2012**; **Ramisch:2015**. Then, we reviewed our data set and the features of the different s gathered. As a result of this study, we revised the taxonomy and modified it to make it conform with the Spanish language.

The remainder of this article is structured as follows: §2 summarizes existing taxonomies and §3 discusses fixedness tests applicable to Spanish and used in our study. §4 explains the creation of our initial data set of Spanish s. In §5, we present the taxonomy we propose for Spanish s based on the results of our re-

search. We also update the information about our data set, expanded to cover all types of s in our new taxonomy. §6 is devoted to the description of the linguistic properties of each type for Spanish. Finally, §7 summarizes our work.

2 Multiword expression typologies

There seems to be a lack of a commonly used taxonomy of s, both in theoretical linguistics and in . In fact, several taxonomies have been proposed throughout the years. Most of them have focused on English s, but as we will point out later in this section, there also exist other taxonomies based on different languages. While it is not the purpose of this section to discuss all existing taxonomies and assess their applicability to the Spanish language and , we think that a brief overview of the state-of-the-art as regards the classification of s is needed. This will not only illustrate the task at hand – finding an taxonomy suitable for Spanish from an point of view – but it will also illustrate the great existing variety of approaches and perspectives.

2.1 MWE taxonomies in theoretical linguistics

As mentioned earlier, several researchers have worked on the analysis and classification of s from a theoretical linguistics point of view. Some of them, such as **moon1998** worked on specific types of s, while others like **Melcuk:1995** and **FillmoreEtAl1988** addressed more general issues. As mentioned by **moon1998**, there is a lack of agreement as far as the terminology on the topic is concerned and she reported the extended discussions of the problem as proof of it. We will not discuss her work here, as her taxonomy – despite being a reference – only focuses on English fixed expressions and idioms and leaves out other important classes such as compound words because they were beyond the scope of her study.

FillmoreEtAl1988 proposed a typology based on the predictability of a construction with respect to the syntactic rules. They distinguished three classes: *unfamiliar pieces unfamiliarly combined*, *familiar pieces unfamiliarly combined* and *familiar pieces familiarly combined*. While *familiar pieces familiarly combined* are formed following the rules of grammar, they have an idiomatic interpretation. *Familiar pieces unfamiliarly combined* require special syntactic and semantic rules, and *unfamiliar pieces unfamiliarly combined* are unpredictable.

Melcuk:1995, on the other hand, used as their criterion the relevance of an expression as a dictionary entry. Their taxonomy is thus mainly based on the se-

mantics of s, and they distinguished between *complete phrasemes*, *semi-phrasemes* and *quasi-phrasemes*. In their approach, *complete phrasemes* are fully non-compositional and would constitute an independent dictionary entry. *Semi-phrasemes* would be those in which at least one of the elements preserves its meaning, and could be listed in the dictionary entry of the base word of the phraseme. Finally, *quasi-phrasemes* are expressions in which all elements keep their original meaning but their combination adds an extra element of meaning, constituting independent dictionary entries.

2.2 MWE taxonomies in Natural Language Processing

s are not only a topic of interest in theoretical linguistics. In research they constitute a major bottleneck for various applications and tools and thus have also been extensively investigated. **Sag:2002** and **Baldwin2010** proposed taxonomies from the point of view of .

Sag:2002 discuss strategies for processing s in applications and thus proposed a taxonomy mainly based on their syntactic fixedness as this is what needs to be modeled to deal with s in a successful way. Figure 1 summarizes their taxonomy. They first distinguish between *lexicalized* and *institutionalized phrases* and then they further divide lexicalized phrases into *fixed* (e.g. *by and large*), *semi-fixed* and *syntactically flexible*. Semi-fixed s include *non-decomposable idioms* (e.g. *to spill the beans*; *to kick the bucket*), *compound nominals* (e.g. *attorney general*; *car park*) and *proper names* (e.g. *San Francisco*; *Oakland Raiders*). Syntactically-flexible s, on the other hand, include *verb-particle constructions* (e.g. *to look up*; *to break up*), *decomposable idioms* (e.g. *to let the cat out of the bag*; *to sweep under the rug*) and *light verbs* (e.g. *to make a mistake*, *to give a lecture*). According to **Sag:2002**, lexicalized phrases are explicitly encoded in the lexicon, whereas institutionalized phrases are only statistically idiomatic.[1]

[**Multiword expressions** [**Lexicalized** [Fixed] [Semi-fixed [Non-Decomposable Idioms] [Compound Nominals] [Proper Names]] [Syntactically-Flexible [Verb-Particle Constructions] [Decomposable Idioms] [Light Verbs]]] [**Institutionalized**, l*=.88]]

Figure 1: Taxonomy proposed by **Sag:2002**.

[1] All examples are taken from **Sag:2002**.

Baldwin2010 carry out a twofold classification. They make a morpho-syntactic classification, and, additionally, they propose an classification based on syntactic variability, which in turn is based on that of **Sag:2002**. In their taxonomy, illustrated in Figure 2, they group compound nominals and proper names into a broader category named *nominal s*. From a morphosyntactic point of view, they distinguish *nominal*, *verbal* and *prepositional s*. Verbal s are further classified into *verb-particle constructions*, *prepositional verbs*,[2] *light-verb constructions* and *verb-noun idiomatic combinations*, and prepositional s are classified into *determinerless-prepositional phrases* (PP-Ds, e.g. *on top*) and *complex prepositions* (complex PPs, e.g. *in addition to*).

[**Multiword expressions**, grow=270 [**Lexicalized**, grow=270 [Fixed, l*=.5
[Non-modifiable
PP-Ds] [Complex
PPs]] [, edge=white] [Semi-fixedP, l*=3.7 [Non-Decomposable
Idioms] [Nominal
MWEs] [PP-Ds with
strict constraints] [Complex PPs]] [, edge=white] [Syntactically-Flexible [,
edge=white] [Verb-Particle
Constructions] [Decomposable
Idioms] [Light
Verbs] [Highly
productive
PP-Ds]]] [**Institutionalized**]]

Figure 2: Taxonomy proposed by **Baldwin2010**.

Ramisch:2012; **Ramisch:2015** proposed a simplified typology based on the morphosyntactic role of the whole in a sentence and its difficulty from an perspective. As illustrated in Figure 3, he identifies three *morphosyntactic classes* (*nominal expressions*, *verbal expressions* and *adverbial and adjectival expressions*) and three additional so-called *difficulty classes* (*fixed expressions*, *idiomatic expressions*, and *"true" collocations*). Nominal expressions are further subdivided into *noun compounds* (e.g. *traffic light*; *Russian roulette*), *proper names* (e.g. *United Nations*, *Alan Turing*) and *multiword terms* (e.g. *profit and loss account*, *myocardial infarction*). Verbal expressions are further subdivided into *phrasal verbs*, which in turn are subdivided into *transitive prepositional verbs* (e.g. *to agree with*, *to rely on*) and *more opaque verb-particle constructions* (e.g. *to give up*, *to take off*); and *light verb constructions* (e.g. *to take a walk*, *to give a talk*).

[2] For **Baldwin2010** *verb-particle constructions* are "a verb and an obligatory particle, typically in the form of an intransitive preposition (e.g. *play around*, *take off*), but including adjectives (e.g. *cut short*, *band together*) and verbs (e.g. *let go*, *let fly*)". *Prepositional verbs* are "a verb and a selected preposition, with the crucial difference that the preposition is transitive (e.g. *refer to*, *look for*)". Although they do not discuss it further, there are cases such as *look forward to*, which would fall into both categories.

[**Multiword expressions**,grow=260 [**Morphosyntactic classes**,grow=289, l*=2.4 [Nominal Expressions, l*=.02 [Nominal compounds] [Proper names, l*=2] [Multiword terms]] [Verbal expressions, l*=3 [Phrasal verbs [Transitive prepositional verbs] [Opaque verb-particle constructions]] [Light verb constructions]] [Adverbial and adjectival expressions, l*=2]] [**Difficulty classes**,l*=.2 [Fixed expressions] [Idioms] ["True" collocations]]]

Figure 3: Simplified taxonomy proposed by **Ramisch:2012**; **Ramisch:2015**.

2.3 Spanish MWE taxonomies

Although Spanish is a widely researched language, few researchers have worked on taxonomies of Spanish s. The main reference for our study could be the seminal work by **Corpas:1996** in Phraseology, who studied Spanish phraseological units, revised previous work and proposed a new taxonomy to classify them. Her taxonomy attempted to establish a classification of Spanish phraseological units based on a set of criteria that should help classify any unit under a specific type. Her taxonomy, summarized in Figure 4, has three major categories subsequently subdivided in more fine-grained subclasses. While *collocations* are classified following their possible part-of-speech patterns (e.g. subject_noun+verb, adjective+noun, etc.), *expressions* are classified according to the syntactic role they may have in a sentence (e.g. *nominal expressions*, *verbal expressions*, *prepositional expressions*...). Finally, *phraseological expressions* are divided into *sentences with a specific value*, *quotes* and *proverbs*.

[**Phraseological Units** [**Phrase** [Grammatically fixed [Collocations]] [Fixed by usage [Expressions]]] [**Sentence** [Fixed by the system [Phraseological Expresions]]]]

Figure 4: Taxonomy of Spanish phraseological units by **Corpas:1996**.

From an point of view, the work by **Corpas:1996** cannot be easily adapted for usage because many classes could be difficult to distinguish from one another. Nominal expressions, for instance, are further subdivided into types following a determined part-of-speech pattern. However, some of these patterns are identical to the ones used to

classify collocations. Thus, to automatically determine whether a "noun+adjective" sequence shall be classified as a collocation (e.g. *enemigo acérrimo* 'archenemy'), or a nominal expression (e.g. *mosquita muerta* 'two-faced person') could be challenging.

Finally, it is also worth mentioning the work by **Leoni:2014**, who also attempted to propose a typology of phraseological units based on the lexical status and the syntactic phenomena of s. In his taxonomy, he first distinguishes between *multi-member lexical units*, which are "units of meaning without necessarily being lexical units", and *collocations*, which are "a lexical choice probably motivated by communication style, with no semantic implications". *Multi-member lexical units* are further divided into lexicalized units (*multi-member lexemes*) and non-lexicalized ones. According to **Leoni:2014**, multi-member lexemes can be characterized by the procedures used to create them. Thus, he distinguishes between those undergoing morphological procedures (*poly-lexemic lexemes*), and those undergoing syntactic procedures (*combined lexemes*). Non-lexicalized units can either be *phrasemes* or *thematic fusions*. He defines *thematic fusions* as "the result of the combination of a supporting verb and a predicative nominal", and *phrasemes* as "unit(s) of meaning formed from at least two open-class lexical morphemes, one of which constitutes the nucleus of the unit and bears the category V". As far as *phrasemes* are concerned, he distinguishes between "continuous expressions that extend across a sentence" (*complete phrasemes*), and "discontinuous expressions that can be replaced by a verb"(*syntagmatic phrasemes*). Figure 5 illustrates his taxonomy.

for tree= fairly nice empty nodes [Poly-lexicality [Multi-member lexical units [Multi-member lexemes [Poly-lexemic] [Combined]] [, l*=.5 [Thematic fusions] [Phrasemes [Complete] [Syntagmatic]]]] [Collocations]]

Figure 5: Taxonomy of Spanish phraseological units by **Leoni:2014**.

In this article, we use the taxonomy proposed by **Ramisch:2012**; **Ramisch:2015** as a starting point for a taxonomy of Spanish s and we combine it with the approach taken by **Sag:2002** and **Baldwin2010** based on syntactic flexibility. This decision was made because these two taxonomies are widely spread among the research community and we wanted to test whether an English-driven taxonomy could be applied to the Spanish language.

3 MWE fixedness tests for Spanish

As one of our objectives was to classify s according to their degree of syntactic flexibility, it is important to determine how this flexibility is going to be measured. Here, we will consider *fixed expressions* those which admit no alteration of their form. *Semi-fixed expressions* will be those which have a certain degree of morphosyntactic variability. This variability, however, is due to the need to conform with the grammatical and orthographical rules of the Spanish language and thus is controlled to a certain extent. From an point of view, these expressions could be easily processed. In the case of fixed s, the

words-with-spaces approach proposed by **Sag:2002** could be used, while in the case of semi-fixed s, this approach could be used adding pointers to the inflected parts of the , just as **Sag:2002** also propose. Finally, flexible s will be those presenting a high degree of variability in their usage (e.g. non-contiguousness, free slots, etc.), which makes their form difficult to predict.

Based on previous work by **Nunberg1994**, where they try to determine the fixedness of s, we designed a set of potential tests to establish the degree of flexibility of Spanish s. This list may be expanded upon further research and, as pointed out by Laporte (this volume), it needs further testing to be supported with statistics. However, we believe that it is a valid starting point for any work on the flexibility of Spanish s and their further linguistic description.

3.1 Inflection

Spanish is a rich morphological language. Thus, the first test that can be used to determine whether an has some degree of flexibility is to check its inflection. In the case of nouns and adjectives, whether or not these can be inflected for number, and in some cases for gender, shall be checked. Generally, adjectives agree in number and gender with the nouns they complement. Thus, their inflection will be dependent on the possibility to inflect their head noun. Examples (1a)–(1b), (2a)–(2b) and (3a)–(3d)[3] exemplify this.

2 anillo de compromiso
N.MASC.**SG** PREP N.MASC.SG
ring of engagement
'engagement ring'

anillo**s** de compromiso
N.MASC.**PL** PREP N.MASC.SG
rings of engagement
'engagement ring**s**' 2 raíz cuadrada
N.FEM.**SG** ADJ.FEM.**SG**
root square
'square root'

raí**ces** cuadrada**s**
N.FEM.**PL** ADJ.FEM.**PL**
roots square
'square root**s**'
lob**o** con piel de cordero
N.**MASC.SG** PREP N.FEM.SG PREP N.MASC.SG
wolf.MASC.SG with skin of lamb
'wolf.MASC.SG in sheep's clothing' lob**a** con piel de cordero

[3] All abbreviations used in this article are listed after the bibliography.

N.**FEM.SG** PREP N.FEM.SG PREP N.MASC.SG
wolf.FEM.SG with skin of lamb
'wolf.FEM.SG in sheep's clothing' lob**os** con piel de cordero
N.**MASC.PL** PREP N.FEM.SG PREP N.MASC.SG
wolves.MASC.PL with skin of lamb
'wolves.MASC.PL in sheep's clothing' lob**as** con piel de cordero
N.**FEM.PL** PREP N.FEM.SG PREP N.MASC.SG
wolves.FEM.PL with skin of lamb
'wolves.FEM.PL in sheep's clothing'

When the includes a pronominal reference to a person, this can also have some variance to agree with the reference. Additionally, when the includes a verb, this can also be inflected for person, tense and mode. Examples (4a)–(4d) and (5a)–(5c), respectively, exemplify this.

el que cort**a** el bacalao
DET.**MASC.SG** PRON.**MASC.SG** V.**3RD.SG**.PRES.IND DET.MASC.SG N.MASC.SG
the who cuts the cod
'big fish.MASC.SG'

la que cort**a** el bacalao
DET.**FEM.SG** PRON.**FEM.SG** V.**3RD.SG**.PRES.IND DET.MASC.SG N.MASC.SG
the who cuts the cod
'big fish.FEM.SG'

los que cort**an** el bacalao
DET.**MASC.PL** PRON.**MASC.PL** V.**3RD.PL**.PRES.IND DET.MASC.SG N.MASC.SG
the who cut the cod
'big fishes.MASC.PL'

las que cort**an** el bacalao
DET.**FEM.PL** PRON.**FEM.PL** V.**3RD.PL**.PRES.IND DET.MASC.SG N.MASC.SG
the who cut the cod
'big fishes.FEM.PL'

Viv**es** a cuerpo de rey.
V.**2ND.SG.PRES.IND** PREP N.MASC.SG PREP N.MASC.SG
live.you by body of king
'You live high on the hog.' Viv**ieron** a cuerpo de rey.
V.**3RD.PL.PAST.IND** PREP N.MASC.SG PREP N.MASC.SG
lived.they by body of king
'They lived high on the hog.' **Hubiera vivido** a cuerpo de rey.
V.1ST/3RD.SG.PAST.SUBJ PREP N.MASC.SG PREP N.MASC.SG
would have lived.I/he/she by body of king
'I/he/she would have lived high on the hog.'

As the variation of this type of s is controlled, in our study all s which only undergo inflection are classified as semi-flexible s.

3.2 Change of determiner

In some cases, the determiner appearing in an is flexible in the sense that there are several items that can occupy that spot within the . Examples (6a)–(6c) illustrate some of the variation of two of the s in our data set.

Nos hicimos **varias** fotos.
PRON.1.PL V.1.PL.PAST.IND ADJ.FEM.PL N.FEM.PL
Ourselves took.1ST.PL several pictures
'We took several pictures.' Nos hicimos **muchas** fotos.
PRON.1.PL V.1.PL.PAST.IND ADJ.FEM.PL N.FEM.PL
Ourselves took.1ST.PL many pictures
'We took many pictures.' Nos hicimos **una** foto.
PRON.1.PL V.1.PL.PAST.IND ADJ.FEM.SG N.FEM.SG
Ourselves took.1ST.PL a picture
'We took a picture.'

In our study, if an *only* undergoes a change of determiner, it is classified as a semi-flexible because this feature can be modeled computationally.

3.3 Pronominalisation

Another useful test to check the degree of flexibility of an is to test whether part of it can be pronominalized. This is only possible for the Noun Phrase and Complementizer Phrase parts of verbal s. Examples (7) and (8) illustrate such cases.[4]

Habíamos quedado para **hacer las fotos** el lunes, pero al final **las** hicimos el martes.
Agreed.to.meet.1ST.PL to make the pictures the Monday, but in.the end them made.1ST.PL the Tuesday
'We had agreed to **take the pictures** on Monday, but in the end we took **them** on Tuesday.'
Después de cenar **dimos un** largo **paseo** por el campo y **lo** disfrutamos mucho.
After of dinner went.1ST.PL a long walk through the field and it enjoyed.1ST.PL a lot
'We **went for a** long **walk** through the field after dinner and we enjoyed **it** greatly.'

When part of a Spanish can be pronominalized, we classify such as a flexible because the fact that not all lexical elements are together in the same clause makes its identification and processing more difficult. While in example (7) the object of the (*las fotos* 'the pictures') is pronominalized and the same verb is used in the second occurrence of the , in example (8) the object is used as the object of a different verb (*disfrutar* 'to enjoy').

3.4 Topicalization

In some cases, it is possible to alter the order in which the elements of an appear. Similarly to what happens with the pronominalisation of s, topicalization is only possible for the Noun Phrase and Complementizer Phrase parts of verbal s. Example (9) shows how

[4] From here on, we omit the morphological analysis of the examples as it is not needed to illustrate the flexibility issues described.

the prepositional phrase (*de política* 'about politics') of a verb with a governed prepositional phrase (*hablar de* 'talk about') may be fronted and appear before the verb itself. Example (10) illustrates how in interrogative sentences the noun phrase of a light verb construction (*qué trato* 'what deal') may also be placed prior to the verb it refers to (*harán*, 'make').[5]

De política no **hablaban** nada más que los domingos.
About politics not talked.3RD nothing more than the Sundays
'They only talked about politics on Sundays.'

¿**Qué trato** crees que **harán** las empresas?
What deal think.2ND that will make.3RD the companies
'What deal do you think the companies will make?'

When an allows for the topicalization of part of it, we classify it as a flexible . An additional reason is that when topicalization occurs, the appears separated in the clause. As it is not possible to determine how many other phrases (and of which type) can appear between the elements of the , its successful processing requires more than just a morphosyntactic analysis.

3.5 Subordinate clauses

s can also appear in complex sentences which have subordinate clauses. In this case, two phenomena may occur. First, the can be partially embedded in a subordinate clause because the element appearing outside of the subordinate clause is also the antecedent of the subordinating conjunction. Example (11) shows this: *el trato* 'the deal' is the antecedent of the subordinating conjunction *que* 'that/which'.

El trato que hizo mi hermana consistía en …
The deal that made my sister consisted in …
'**The deal** my sister **made** involved …'

Second, part of the can be the antecedent of a subordinate clause, as in (12).

Mi hermana **hizo un trato que** consistía en …
My sister made a deal that consisted in …
'My sister **made a deal that** involved …'

When a part of an can be embedded in a relative clause or be the antecedent of a relative clause, we classify it as a flexible .

3.6 Passivization

A frequent way of testing the flexibility of English s is to test whether or not their passivization is possible. As the passive voice is not as frequent in Spanish as in English, this test may not be very informative for testing Spanish s. Moreover, in Spanish there are two passivization mechanisms:

[5] In this example, a second phenomenon occurs, as the verb is part of a subordinate clause whereas the noun phrase is part of the main clause. This is discussed in the next flexibility test in §3.5.

1. Passives using the auxiliary verb *ser* 'to be'; and
2. Passives using the pronoun *se*, also called 'passive *se*'.

Passives using the auxiliary verb *ser* are not very frequent, and it is common to find 'passive *se*' sentences.

In the case of s, this test can still be used, and in some cases, such as the one in example (13), it will be possible to find an appearing in a passive voice construction. In some cases, both types of passives are possible. Example (14), shows how the passivization of example (13) could be also done by means of the Spanish pronoun *se*.

La decisión fue tomada el lunes.
The decision was taken the Monday
'**The decision was made** on Monday.' **La decisión se tomará** el lunes.
The decision itself will be taken.3RD.SG the Monday
'The decision will be made on Monday.'

If an can *only* undergo passivization (i.e. all other tests are negative), we classified it as semi-flexible. Else, we classified it as a flexible .

3.7 Appearance of other elements

In some cases, other elements such as adjectives, adverbs or pronouns which do not belong to the appear embedded in the . The number of elements that can appear embedded in the also varies. There could be only one element, or several. Examples (15) to (17) illustrate this.

dar un **largo** paseo
to take a long walk
'to take a **long** walk' dar un **largo y agradable** paseo
to take a long and nice walk
'to take a **long and nice** walk' echar **profundamente** la siesta
to take deeply the nap
'to take a nap **deeply**'

When other elements can appear embedded within the elements of an we classified it as a flexible .

3.8 Ellipsis

Finally, part of an can sometimes be omitted. This is usually the case when, for instance, the object of an has been mentioned earlier and then it is referred to at a later stage. Example (18) illustrates this. In the example, the complement of the verb *hacer* 'to do' is elided but *qué* 'what' is used to refer to it 'what deal'.

¿**Qué** crees que **harán**?
What think.2ND.SG that do.3RD.PL
'**What (deal)** do you think they will **do**?'

Ellipsis may also occur when there is coordination. Example (19) illustrates this by showing two coordinated main clauses that share the same predicate (*quedarse* 'to keep

for oneself') with a change both of the subject (*María–Juan*), and of the complement of the prepositional phrase governed by the verb (*el libro* 'the book' vs. *el disco* 'the disc').

María **se quedó con** el libro y Juan **con** el disco.
María herself kept with the book and Juan with the disc
'María kept the book and Juan the disc.'

In those cases in which an allows for the omission of part of it, we classified the as a flexible .

4 Creating a data set to analyze Spanish MWEs

As a starting point for our study, we took the MWE taxonomy proposed by **Ramisch:2012**; **Ramisch:2015** and created a preliminary data set of Spanish s. It was not compiled by doing a corpus analysis and subsequently trying to analyze and classify the s detected, but rather by taking the English examples from **Ramisch:2012**; **Ramisch:2015** and trying to find similar ones in Spanish. The preliminary data set consisted of 150 Spanish s classified according to Ramisch's taxonomy (**Parra:2015**).

Figure 6 exemplifies all of the types distinguished in Ramisch's taxonomy with Spanish examples and their translations into English. As may also be observed, there is no example for *phrasal verbs*. This is because Spanish lacks such a type of , although there are verbs with a governed prepositional phrase (e.g. *acordarse de* 'to remember') which, to a certain extent, have a similar behavior to that of English phrasal verbs.[6]

We then analyzed and classified the s by their degree of difficulty for purposes. To this aim, we used the "fixed, semi-fixed, flexible" classification proposed in the papers by **Sag:2002** and **Baldwin2010**.

[6] As pointed out in the annotation guidelines for the PARSEME shared task on automatic detection of verbal multiword expressions (**Vincze:2016**), Verb Particle Constructions (also called phrasal verbs), "are pervasive in English, German, Hungarian and possible other languages but irrelevant to or very rare in Romance and Slavic languages or in Farsi and Greek for instance". As **Vincze:2016** also point out, contrary to inherently prepositional verbs (referred to in this paper as *verbs with a governed prepositional phrase*), the particle present in phrasal verbs cannot introduce a complement.

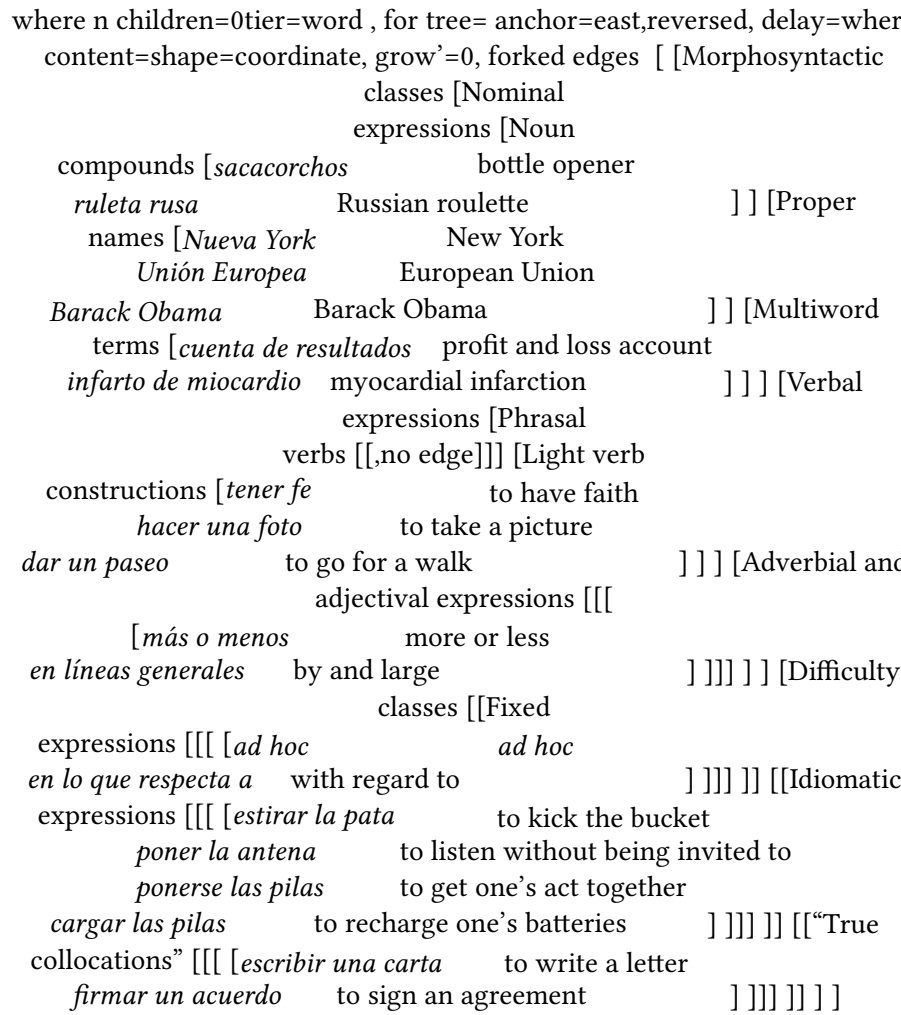

Figure 6: Spanish MWEs classified following **Ramisch**:**2012**; **Ramisch**:**2015** taxonomy.

The Spanish Grammar[7] (**RAE:2010**) was also used to detect additional types not present in the taxonomy, describe subclasses, and gather further examples for our data set. As we aimed at having a number of entries for each type that allowed us to properly describe its features, additional new entries were also added to the data set. Appendices 7, 7 and 7 comprise our data set classified in fixed, semi-fixed and flexible s respectively.

5 Our Spanish MWE taxonomy

When creating our data set, we realized that the taxonomy we had started to work with was not completely matching the Spanish s we were gathering. Thus, we started to modify the taxonomy and adapt it to the Spanish language. This confirms the common criticism against current taxonomies claiming they are based on the English language and that other languages cannot be classified in the same way.

After revising our data and discussing the different categories we had encountered, we first decided to eliminate the types *compound nouns* and *multiword terms* and add a new category, *complex nominals*, to account for single-token compound nouns in Spanish such as *abrebotellas* 'bottle opener', and syntagmatic compounds such as *botella de vino* 'wine bottle'.

The concept of complex nominals was already introduced by **Atkins:2001** to account for complex nominal constructions in languages other than English that can be considered s. While compounds in Germanic languages such as English or German are created by appending several nouns together in either several tokens (e.g. English) or one (e.g. German, Norwegian), in Spanish (and other Romance languages such as Italian or French), these expressions require the usage of prepositions and articles and show a different structure.

Multiword terms were eliminated as an type in our taxonomy because the different types of terms could be actually classified within other types in our taxonomy. Terms might be either single words (e.g. *fideicomiso* 'trust') or more complex structures, ranging from complex nominals (e.g. *cuenta de resultados* 'profit and loss account') to verbal s (e.g. *fallar a favor* 'to rule in favor') and idiomatic s (e.g. *a tenor de lo dispuesto en* 'in accordance with/under the stipulations of'), which justified their reclassification into other categories in our new taxonomy. Moreover, terminology is a different research field with its own taxonomies for classifying terms. The terms gathered in our data were thus redistributed in the other types in our taxonomy.

Adjectival and adverbial s had to be split in two different categories as they do not share the same features. Moreover, a closer look at *adjectival expressions* revealed that in Spanish we can distinguish between three different main subclasses: *compounds*, *adjectival phrases* and *adjectives with a governed prepositional phrase*.

In the case of *verbal expressions*, we deleted *phrasal verbs* because, as explained earlier (c.f. §4), Spanish does not have such type of verbs. In order to cover other types in

[7] In this article, we use italics to refer to the Spanish grammar written by the *Real Academia de la Lengua Española* (RAE, Royal Spanish Language Academy) used as a reference in our work.

Spanish, we had to add three new subclasses: *periphrastic constructions*, *verbal phrases* and *verbs with a prepositional phrase*.

We also decided to eliminate the *fixed expressions* from the taxonomy as this refers to a type of flexibility rather than a type of . According to **Ramisch:2015**, "they correspond to the fixed expressions of **Sag:2002**, that is, it is possible to deal with them using the words-with-spaces approach. Such expressions often play the role of functional words (*in short*, *with respect to*), contain foreign words (*ad infinitum*, *déjà vu*) or breach standard grammatical rules (*by and large*, *kingdom come*)". The fixed expressions present in our data set could easily be redistributed across two additional types added to the morphosyntactic types: *conjunctional phrases* and *prepositional phrases*. Foreign s have been excluded of our study because their classification and characterization is beyond the scope of this article.

As far as the other two *"difficulty classes"* in the taxonomy proposed by **Ramisch:2012**; **Ramisch:2015**, we also eliminated them as they did not comply with our aim of classifying s by morphosyntactic types and rather constituted categories based on semantic criteria (*idioms*), or statistical co-occurrence (*"true" collocations*). We reclassified all items in those categories across several of the morpho-syntactic types: *complex nominals*, *light verb constructions* and *verbal phrases*. To accommodate the remaining few items that could not be reclassified, we created a new and broader category: *sentential expressions*.

Our taxonomy comprises two different axes: *morphosyntactic type* and *flexibility degree*. The *morphosyntactic type* axis is based on Ramisch's (**Ramisch:2012**; **Ramisch:2015**) taxonomy with the modifications explained above. The *flexibility degree* axis is based on the three levels of flexibility identified by **Sag:2002** and **Baldwin2010**. Thus, all s in our data set are classified according to their morphosyntactic type and flexibility.

Figure 7 shows our taxonomy and its two main axes: the type and the flexibility degree. It also quantifies the number of samples in our data set per morphosyntactic type and flexibility.

6 The linguistic properties of Spanish MWEs

In what follows we analyze the Spanish s in our data set per type and describe their main linguistic properties. The analysis was carried out manually and complemented by making searches in Spanish written corpora when we needed to verify our linguistic intuition of a particular .[8] Specifically, we used two contemporary Spanish corpora: CREA[9] and CORPES XXI[10].

All entries in our data set were manually analyzed.[11] Our manual study, combined

[8] A deeper corpus study of the MWEs gathered in our data is planned as future work.

[9] Corpus de referencia del español actual (Reference Corpus for Current Spanish): http://corpus.rae.es/creanet.html.

[10] Corpus del español del siglo XXI (Corpus for 21st Century Spanish): http://web.frl.es/CORPES/view/inicioExterno.view.

[11] As mentioned earlier, the inflectional morphology of Spanish is richer than the morphology of English and therefore it requires a more detailed linguistic analysis. A similar observation was made in **Savary:2008** and **Gralinski:2010**, who studied the complexity of encoding MWEs

2*			Flexibility degree		
			Fixed	Semi-fixed	Flexible
14*Morphosyntactic types	3*Adjectival expressions	Adjectival compounds	—	10	—
		Adjectival phrases	14	2	2
		Adjectives with a governed prepositional phrase	—	—	13
	Adverbial expressions		49	1	1
	Conjunctional phrases		10	—	—
	3*Nominal expressions	Complex nominals	23	43	—
		Proper names	35	—	—
		Nouns with a governed prepositional phrase	—	—	12
	Prepositional phrases		10	—	—
	4*Verbal expressions	Light verb constructions	—	—	42
		Periphrastic constructions	—	—	19
		Verbal phrases	—	11	15
		Verbs with a governed prepositional phrase	—	—	21
	Sentential expressions		4	1	—

Figure 7: New MWE taxonomy for Spanish.

with the grammar study and the corpus queries, allowed us to identify and verify the specific linguistic features of Spanish s described here.

6.1 Adjectival expressions

6.1.1 Adjectival compounds

Adjectival compounds in Spanish are one typographic word (e.g. *drogadicto* 'drug addicted'; *pelirrojo* 'redheaded'). They are usually formed by joining two adjectives together, or a noun and an adjective. Although they constitute one typographic word, we consider them multiwords because they are composed of several words and might need to be processed in a special way in some applications (like Machine Translation), as German compounds, for instance.

In our data set, all adjectival compounds are semi-flexible.[12] They inflect either in gender (masculine/feminine) and number (singular/plural), or only in number (singular/plural).[13] In some cases, these adjectival compounds are nominalized in usage, despite

in morphologically rich languages such as Polish and French. Testing the formalisms they propose is beyond the scope of this article.

[12] C.f. Figure 7.

[13] See Appendix 7.

them being adjectives. For instance, *drogadicto* can occur in a sentence as an adjective or a nominalized adjective. Examples (20) and (21) illustrate this.

Ella está ayudando a un hombre drogadicto.
She is helping to a man.N drug.addicted.ADJ
'She is helping a drug addicted man.' Ella está ayudando a un drogadicto.
She is helping to a drug.addicted.N
'She is helping a drug addict.'

6.1.2 Adjectival phrases

According to *the Spanish Grammar* (**RAE:2010**), adjectival phrases are lexicalized phrases that behave syntactically like adjectives. Many have the structure of a prepositional phrase which complements a head noun, and sometimes are equivalent to adverbial collocations complementing predicates (e.g. *juramento* ***en falso*** 'a lie under oath' vs. *jurar* ***en falso*** 'to lie under oath'). Alternatively, they can also be of the form *como* 'as' followed by a nominal phrase (e.g. *como una catedral* 'huge'). Finally, it is also possible to find adjectival phrases formed by adjectives in coordination (e.g. *corriente y moliente* 'plain ordinary').

The majority of the adjectival phrases gathered in our data set are fixed (14), although we also registered 2 semi-fixed phrases and 2 flexible ones. The 2 flexible phrases are of the type "preposition + noun", whereas in the semi-fixed ones one has the Part-of-Speech () pattern "preposition + adjective + noun" and the other one is of the type "adjective + conjunction + adjective". Moreover, all these patterns are also present among the 14 fixed ones, which suggests that there is not a preferred form that flavors flexibility.[14] This seems to be in line with the fact that these phrases are lexicalized, and thus show a tendency to be invariable.

6.1.3 Adjectives with a governed prepositional phrase

Adjectives with a governed prepositional phrase are adjectives that are always followed by a certain preposition. The preposition is not predictable, since it is due to both semantic and historical reasons. Moreover, in some cases the prepositional phrase has to be explicit (e.g. *carente de* 'deprived of'), whereas in other cases where the information is considered to be implicit, the prepositional phrase can be omitted (e.g. *ser fiel a* 'to be loyal to').

We gathered 13 adjectives with a governed prepositional phrase in our data set. All of them are fully flexible, as they can be modified not only according to number (singular/plural) and gender (masculine/feminine), but also allow for other elements such as adverbs to be inserted between the adjective and the prepositional phrase.

[14] This shall however be confirmed by undergoing a corpus based analysis of all items in our data set and new ones.

6.2 Adverbial expressions

According to *the Spanish Grammar* (**RAE:2010**), adverbial expressions are fixed expressions formed by several words that account for a single adverb. They might not have the form of an adverb, but they function as such. Some can be substituted by adverbs ending in *-mente* (e.g. *en secreto* 'in secret' and *secretamente* 'secretly'), but most of them have a more specific or slightly different meaning from the adverbs which are morphologically similar to the adverbial expression.

There are some very exceptional cases in Spanish in which adverbial expressions can be slightly modified (**RAE:2010**) by adding a suffix to the main noun (e.g. *a golpes/a golpetazos*,[15] 'violently'; lit. 'by hits/by thumps') or introducing an adjective between two elements of the expression (e.g. *a mi entender/a mi modesto entender* 'by my understanding/by my modest understanding').

There are three different types of adverbial expressions in Spanish:

- "Preposition + noun phrase", where the noun phrase may be a single noun (e.g. *por descontado* 'of course'), or a noun modified by other elements such as determiners or adjectives (e.g. *a la fuerza* 'by force');
- "preposition + adjective/participle" (e.g. *a escondidas* 'behind somebody's back'; *por supuesto* 'of course'); and
- "lexicalized phrase" which typically expresses quantity, manner and/or degree (e.g. *una barbaridad* 'quite a lot'; *codo con codo* 'elbow to elbow').

We gathered a total of 51 adverbial expressions in our data set. 28 of them are of the type "preposition + noun phrase" (12 in which the noun phrase is a single noun and 16 in which the noun phrase includes modifiers); 11 are of the type "preposition + adjective/participle", and the remaining 12 are lexicalized phrases expressing quantity, manner or degree. A manual analysis of these 51 items revealed that adverbial expressions in Spanish are mostly fixed in their structure, which confirms what is stated in *the Spanish Grammar* (**RAE:2010**).

6.3 Conjunctional phrases

Conjunctional phrases are groups of words containing a conjunction that function as a single conjunction (e.g. *a fin de que* 'in order to'). In Spanish, once identified, this type of s is easy to deal with from an perspective. They are invariable and do not allow the inflection of any of its parts, which would allow to process them successfully using the words-with-spaces approach used with other fixed expressions. 10 conjunctional phrases were included in our data set.

[15] In Spanish, the suffix *-azo* is a very productive suffix with different meanings. Here, it is used as an augmentative to indicate the size or strength of the blow.

6.4 Nominal expressions

6.4.1 Complex nominals

We have defined this category similarly to what **Atkins:2001** propose. Thus, it accounts for noun compounds in Spanish, and includes other nominal phrases that usually behave as nominal compounds in other languages such as English. *The Spanish Grammar* (**RAE:2010**) accounts for several types of compounds in Spanish:

- **Noun compounds of one typographic word**: *cascanueces* 'nutcracker', *limpiacristales* 'window cleaner', *aguafiestas* 'spoilsport'.
- **Noun compounds of two typographic words**: two nouns after one another as in *mesa camilla* 'round table', *hombre lobo* 'werewolf'; or a noun followed by an adjective as in *guerra civil* 'civil war'.
- **Syntagmatic compounds**: nominal phrases typically including a prepositional phrase as in *goma de borrar* 'eraser', *café con leche* 'coffee with milk', *el día a día* 'everyday life', *ley de la jungla* 'law of the jungle'.

We gathered a total of 66 complex nominals in our data set. A manual analysis of these 66 items revealed that complex nominals in Spanish are either fixed in their structure (23), or semi-fixed (43).

We further classified our data according to the three types described above. 11 items were noun compounds of one typographic word, 19 items were noun compounds of two typographic words, and the rest (36) were syntagmatic compounds. All compounds of one typographic word in our data but one are fixed and do not experience any kind of morphosyntactic variation in their usage. However, this does not hold true for all Spanish noun compounds of one typographic word. In our data, most of the noun compounds we gathered end in *-s*, which means that both the singular and the plural forms of such noun compounds are the same. Other noun compounds, such as the only one we gathered as semi-fixed, *bocacalle* 'side-street' do inflect in plural (*bocacalle**s***).

19 items were noun compounds of two typographic words. In 2 cases these noun compounds are fixed and do not show any kind of variance: *vergüenza ajena* 'the feeling of being embarrassed for somebody', and *gripe aviar* 'avian influenza'. The remaining items can be inflected in either singular or plural and thus are semi-fixed. We gathered 13 items of the type "noun + adjective", and 6 of the type "noun + noun". While the compounds of the type "noun + adjective" seem to require that both the noun and the adjective are inflected and agree in number, in the case of the "noun + noun" compounds this does not always hold true. In some cases, only the head of the compound can be inflected in the plural forms (e.g. *ciudad dormitorio* 'dormitory town' vs. *ciudad**es** dormitorio* 'dormitory towns'; and *niño prodigio* 'child prodigy' vs. *niño**s** prodigio* 'child prodigies'). *The Spanish Grammar* (**RAE:2010**) points out that when the modifier of the compound adopts an adjectival function (e.g. *disco pirata* 'pirated CD'; *momento clave* 'key moment'), the plural form of the compound can be formed by only inflecting the head of the compound[16] (e.g.

[16] In Spanish, the head of a compound is the left-most element in the compound.

discos pirata 'pirated CDs'; *momentos clave* 'key moments') or both nouns, the head and the modifier (e.g. *discos piratas*; *momentos claves*).

Finally, the remaining 36 items in our data set were *syntagmatic compounds*. 11 of them are fixed, while the other 25 are semi-fixed.

Complex nominals in Spanish can only inflect in terms of number. Although there seems to be a pattern in which only the head of the compound is inflected (e.g. *ciudad/ciudades dormitorio* 'dormitory town/towns'), it is not always the case.

For purposes, an easy strategy to test whether a complex nominal is fixed or allows for inflection would be to inflect the complex nominal in number and check whether that form can be found in a monolingual corpus. If it is not the case, the complex nominal is fixed. Otherwise, it is semi-fixed.

6.4.2 Proper names

Proper names identify a being among others without providing information of its features or its constituent parts. These nouns do not express what things are, but what their name is as individual entities. Proper names have referring capacity, do not participate in lexical relations and, strictly speaking, cannot be translated (*Spanish Grammar* **RAE:2010**).

The Spanish Grammar (**RAE:2010**) identifies two types of proper names: anthroponyms and toponyms. However, it also argues that names that account for festivals or celebrations, celestial bodies, allegorical representations, works of art, foundations, religious orders, companies, clubs, corporations and other institutions share the same characteristics.

We gathered a total of 35 proper names in our data set. A manual analysis of these 35 items revealed that proper names in Spanish cannot be morphologically modified.

We classified our data according to the three types listed above. 12 items were toponyms, 11 items were anthroponyms, and 12 were classified under "others", which include celestial bodies, works of art, foundations, companies, clubs, corporations, etc. All those items do not have any kind of morphological variation.

6.4.3 Nouns with a governed prepositional phrase

Nouns with a governed prepositional phrase are nouns that are always followed by a certain preposition. Occasionally, more than one preposition is possible (e.g. *actitud con/hacia/respecto de* 'attitude with/towards/regarding'). This is usually the case when the phrase following the preposition indicates matter, direction or addressee. In some cases, two prepositions with exactly the same meaning are valid (e.g. *asalto a/de* 'assault to/on'; *solución a/de* 'solution to/of').

Some nouns followed by a prepositional phrase derive from the verbal form, maintaining the same preposition (e.g. *oler a*/*olor a* 'to smell like'/'smell of'; *eximir de*/*exento de* 'to exempt from'/'exempt from'). There are cases, though, where the preposition changes (e.g. *amenazar con*/*amenaza de* 'to threaten to'/'threat of'; *interesarse por*/*interesado en* 'to be interested in'/'interested in').

We gathered 12 nouns with a governed prepositional phrase. As the adjectives with a governed prepositional phrase, all of them are fully flexible. They can be modified according to number (singular/plural) and gender (masculine/feminine), and they admit an adverb and/or an adjective between the noun and the preposition.

6.5 Prepositional phrases

Prepositional phrases are groups of words containing a preposition that function as a single preposition (e.g. *en detrimento de* 'at the expense of'). Similarly to conjunctional phrases (c.f. §6.3), these s are fixed in Spanish and thus none of its parts can inflect. Our data set includes 10 prepositional phrases.

6.6 Verbal expressions

6.6.1 Light verb constructions

Light verb constructions () in Spanish are semi-lexicalized verb constructions formed by a verb with a supporting role or semantically weak complemented by an abstract noun[17] (**RAE:2010**). *The Spanish Grammar* (**RAE:2010**) identifies the following light verbs in Spanish: *dar*, 'to give'; *tener*, 'to have'; *tomar*, 'to take'; *hacer*, 'to do' or 'to make', and *echar*, 'to throw'. In some cases, the noun is preceded by an article. Many s can be paraphrased using another single verb with similar meaning (e.g. *dar un paseo*: *pasear* 'to take a walk': 'to walk'; *hacer alusión*: *aludir* 'to make an allusion': 'to allude').

This definition thus differs from the one offered by Laporte (this volume), as well as with the one specified in the annotation guidelines for the PARSEME shared task on automatic detection of verbal multiword expressions (**Vincze:2016**). **Vincze:2016** identify the following six general characteristics of s:

1. They are formed by a verb and its argument containing a noun. The argument is usually a direct object, but sometimes also a prepositional complement or a subject.
2. Both the verb and the noun (included in the complement) are lexicalized.
3. The verb is "light", i.e. it contributes to the meaning of the whole only to a small degree.
4. The noun has one of its regular meanings.
5. The noun is predicative, and in s one of its arguments becomes also a syntactic argument of the verb. Moreover, the subject is usually an argument of the noun.
6. The noun typically refers to an action or event.

[17] *The Spanish Grammar* (**RAE:2010**) defines abstract nouns as those nouns which refer to something of a non-material nature such as actions, processes and attributes that we assign to beings when we think of them as independent entities (e.g. beauty, dirt).

Bearing in mind that our ultimate goal is to find a taxonomy of Spanish s that can be used from an point of view, we took here a rather comprehensive approach and combined both definitions. Thus, the s in our data set include both expressions including the light verbs identified by *The Spanish Grammar*, and other verbs that in combination with certain nouns can be considered light because their meaning is bleached to a certain extent.

We gathered a total of 42 s in our database. The verbs contained in light verb expressions always inflect in person (1st, 2nd, 3rd / singular or plural), tense (present, past or future) and mode (indicative, subjunctive or imperative), just as any other verb. Most of the times, the other elements of the expression (article and noun) can also be modified without changing the meaning of the expression (e.g. *dar un beso* 'to give a kiss'; *dar dos besos* 'to give two kisses').[18] In our data set, the noun phrases of 10 of the 42 s can appear either in singular or plural. There are some exceptional cases in which the meaning of the expression changes when the noun is singular or plural (e.g. *tener gana*, 'to be hungry' vs. *tener ganas* 'to feel like'; *hacer ilusión* 'to look forward to' vs. *hacerse ilusiones* 'to get one's hope up').[19] Finally, adjectives and adverbs can be included between the different elements of the expression (e.g. *echar* ***profundamente*** *la siesta*, 'to take a nap **deeply**'; *echar una* ***larga*** *siesta*, 'to take a **long** nap'), which means that they are flexible s.

Regarding other flexibility tests such as pronominalisation, topicalization, subordinate clauses and passivization,[20] further research in large Spanish corpora would be required. It seems that most constructions do allow for the pronominalization of the noun (c.f. example (8)) and the appearance of subordinate clauses (e.g. *El paseo que dimos ayer* 'The walk we took yesterday'), while they do not seem so prone to allow for topicalization or passivization.

From an perspective, light verb expressions are challenging in Spanish. While some issues such as the verb tenses can be targeted specifically, some other issues require the usage of other processing strategies. Thus, a change in the determiner or the insertion of adjectives and adverbs between the different elements of the expression will require the design of specific strategies to successfully identify and process these s.

6.6.2 Periphrastic constructions

Verbal periphrastic constructions in Spanish are syntactic combinations in which an auxiliary or semi-auxiliary verb is used in combination with a past participle, an infinitive or a gerund and both verbs constitute a unique predicate (*Spanish Grammar* **RAE:2010**). The verb used as an auxiliary can also appear in non-periphrastic constructions having its full meaning. In some cases, these constructions include the usage of a preposition (e.g. *empezar a …* 'to begin to …'; *acabar de …* 'to have just finished to …').

[18] For more examples of changes in the determiner, see Examples (6a) to (6c).

[19] These cases are registered in our data set as different MWE entries.

[20] C.f. §§3.3–3.6.

The first verb in the periphrastic construction is the one which undergoes inflection, whereas the second one always appears in the same non-finite form, and it is the one which varies and constitutes the main verb of the clause. Sometimes, as example (22) shows, an element such as an adverb can appear between the first element of the periphrasis and the second one. The subject can also appear in between the main verb and the auxiliary or semi-auxiliary verb (example (23)).

Tuvo **casi** que saltar para no caerse.
Had.3RD.SG.MASC/FEM almost that jump for not fall.himself/herself.
'He/she almost had to jump to avoid falling down.'

No podía **yo** creérmelo, pero …
Not could I believe.it, but …
'I could not believe it, but …'

We gathered a total of 19 periphrastic constructions in our data set. Due to their variability in inflection and the allowance of other elements, we have tentatively classified them as flexible. However, further research is needed to determine if certain types could be considered semi-flexible (i.e. those in which the only undergoes inflection) because these structures do not seem to allow for pronominalization, topicalization, subordination or passivization.

Prometió comprar el libro.
Promised.3RD.SG.MASC/FEM buy the book
'He/she promised to buy the book.'

Pudo comprar el libro.
Could.3RD.SG.MASC/FEM buy the book
'He/She could have bought the book.'

One problem of this type of construction is that sometimes it has the same structure as a non-periphrastic one. There are cases, in which a full verb is followed by another verb in a non-finite form, and is the head of the predicate, while the non-finite form is introducing a subordinate clause which complements the main verb. In such cases, there is no periphrasis. In other cases, the same structure ("inflected verb + verb in non-finite form") act as a single unit. In such cases, the inflected verb acts as an auxiliary or semi-auxiliary verb, while the main verb is the one in non-finite form. Examples (24) and (25) illustrate this. In (24), *comprar el libro* 'buy the book' would be a subordinate infinitive clause that is the direct object of the predicate (*prometió* 'promised') of the main clause. In (25), however, *pudo comprar* 'could have bought' is the predicate of the clause and *el libro* 'the book' is its direct object. This makes this type of constructions particularly tricky to detect and to process.[21]

[21] This type of structure is worth researching within a larger project including large corpus searches. This is beyond the scope of this article, where we only aim at detecting MWE types in Spanish that are not covered in the current MWE taxonomies explained in §2.

6.6.3 Verbal phrases

Verbal phrases are those s whose head is a verb and which cannot be classified as any other type of verbal s. All of them share the feature that to a certain extent they are idiomatic expressions whose semantics are non-compositional. As we aimed at classifying Spanish s from a morphosyntactic point of view, many of the items that we originally had classified as idioms following Ramisch's taxonomy (**Ramisch:2012**; **Ramisch:2015**) are classified as verbal phrases in our data set.

In total, 26 items of our data set were classified as verbal phrases. 11 of them were classified as semi-fixed s and the remaining 15 as flexible s. In all the verbal phrases classified as semi-fixed the verb appearing in the inflects (e.g. *coger el toro por los cuernos* 'to take the bull by the horns'; *empezar la casa por el tejado* 'to put the cart before the horse').

Finally, we detected cases in which it was also possible for other words to appear within the to modify its meaning. In these cases, besides the verb inflection and the noun singular/plural and masculine/feminine alternations, the could include other modifying elements. For example, *entrar al trapo* 'to respond to provocations', can be modified by elements referring to its frequency (e.g. *entrar* ***siempre*** *al trapo* 'to respond to provocations **always**').

Another special type of flexibility is the one created by the presence of reflexive pronouns as part of the verb in the , because depending on the overall structure of the sentence the pronoun may appear in different parts of it. Examples (26a) to (26c) below show this phenomenon with the *irse de la lengua* 'to let the cat out of the bag'.

No tienes que irte de la lengua
ADV V.**2ND.SG**.PRES.IND PRON V.INF+PRON.**2ND.SG** PREP DET.FEM.SG N.FEM.SG
not have(.you) that go.yourself of the tongue
'Do not let the cat out of the bag.'
No te tienes que ir de la lengua
ADV PRON.**2ND.SG** V.**2ND.SG**.PRES.IND PRON V.INF PREP DET.FEM.SG N.FEM.SG
not yourself have(.you) that go of the tongue
'Do not let the cat out of the bag.' Prometió que no se iría de la lengua
V.**3RD.SG**.PAST.IND PRON ADV PRON.**3RD.SG** V.3.SG.COND.IND PREP DET.FEM.SG N.FEM.SG
Promised.MASC/FEM that not himself/herself would go of the tongue
'He/she promised not to let the cat out of the bag.'

As s in which a reflexive verb appears also allow for other types of flexibility such as the apparition of modifiers, we classified them as flexible s. However, most of these verbal phrases do not occur undergoing other types of flexibility such as topicalization or passivization and further research is needed to confirm their flexibility degree.

6.6.4 Verbs with a governed prepositional phrase

Verbs with a governed prepositional phrase are verbs that are always followed by a certain preposition.[22] The preposition is not predictable, since it is due to both semantic and historical reasons. Usually, only one preposition governs the phrase, though occasionally more than one is possible, especially in those cases where the phrase following the preposition indicates matter, direction or addressee (e.g. *hablar de/sobre/acerca de* 'to talk of/about'; *viajar a/hacia/hasta* 'to travel to/towards').

Spanish reflexive verbs usually have a governed prepositional phrase (e.g. *arrepentirse de* 'to regret'; *referirse a* 'to refer to'), and a few show a possible alternation between the governed prepositional phrase and a direct object (e.g. *quedarse algo/quedarse con algo* 'to keep something'). Finally, some verbs require a governed prepositional phrase for some of their meanings. In such cases, the meaning of the verb is determined by the occurrence of a governed prepositional phrase (e.g. *entender algo*/*entender de algo* 'to understand something'/'to know about something').

We gathered a total of 21 verbs with a governed prepositional phrase. A manual analysis revealed that the verb can always inflect in terms of person, tense and mode. As other elements may intervene between the verb and the prepositional phrase, and the prepositional phrase can sometimes undergo topicalization (see example (9)), we tentatively classified all of them as flexible.

6.7 Sentential expressions

Some of the s that we included in our data set constitute full clauses. They all share the fact that they are idiomatic expressions as well. However, as we aimed at classifying s from a morphosyntactic point of view, we have classified them as "sentential expressions".

In our data set, only 5 s of this type have been gathered. 4 of them are fixed, whereas 1 is semi-fixed: *la gota que colma el vaso* 'straw that breaks the camel's back'. Their main difference is that while the fixed ones are fully lexicalized (e.g. *cuando el río suena, agua lleva* 'when there is smoke, there is fire'), the semi-fixed allows for verb inflection.

If we consider Spanish proverbs as sentential expressions, this class of our data set could be expanded greatly. However, at this point we do not aim at finding a way of automatically identifying such exceptional cases and characterizing them.[23]

7 Conclusion

In this article, we have analyzed the different types of Spanish s we identified. The starting point of our research was a data set created on the basis of an existing taxonomy for

[22] They are similar in this sense to the adjectives and nouns with a governed prepositional phrase described in Sections 6.1.3 and 6.4.3.

[23] The *Centro Virtual Cervantes* (Instituto Cervantes), has a collection of Spanish proverbs translated to other languages and with useful information about their variants and synonyms that could be used for further research (http://cvc.cervantes.es/lengua/refranero/Default.aspx).

s. Upon our linguistic analysis, we realized that such taxonomy was not adequate for describing Spanish s and we modified it to accommodate our findings.

One interesting finding is the fact that in Spanish there seem to be some categories that are only fixed (*conjunctional phrases*, *prepositional phrases* and *proper names*), or only flexible (*light verb constructions*, *adjectives, nouns and verbs with governed prepositional phrases* and *verbal periphrastic constructions*). Only *adjectival compound*s are exclusively semi-flexible. The other types having semi-flexible s are either also fixed (*complex nominals* and *sentential expressions*), also flexible (*verbal phrases*) or both fixed and flexible (*adjectival expressions* and *adverbial phrases*).

It also seems clear that typologies should be adapted to the language under research, and classic typologies mainly based on the English language do not seem adequate to describe and classify s in other languages. Our research is proof of this fact. Moreover, the taxonomy proposed here has also shown ways of integrating the traditionally considered "difficulty class" of *idioms* within the morphosyntactic classes.

We believe that our work is novel in the sense that we have tested an existing taxonomy to classify Spanish s. In future work we intend to validate our data set asking other linguists whether they agree or not with our classification. We also intend to expand it for the categories underrepresented and carry out further corpus searches to validate our analyses.

Another possible path to explore would be to evaluate the extent to which the flexibility tests discussed in §3 are valid and whether specific types of s require specific tests. It would also be interesting to explore the word-span between the different parts of s and whether discontinuous s in Spanish share some features. This would enable their automatic identification and processing in applications.

From a multilingual perspective, it would be interesting to further compare our data set with the translations of its entries into other languages. This is interesting from a traductological point of view, as it would allow to further compare s and their behavior in different languages. Our data set includes the translations into English of all the items. Many Spanish s translate as English s. In fields such as translation studies or Machine Translation, a further study of these correspondences would be highly relevant.

Finally, it would also be interesting to see if language families share a common taxonomy. We have argued here the need of a language-specific taxonomy. However, it could be that languages belonging to the same language family share a taxonomy and thus instead of language-specific taxonomies there is a need for language-family specific taxonomies.

Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable feedback.

Carla Parra Escartín was supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n° 317471.

Abbreviations

.45lQ 1/2/3 first/second/third person
ADJ adjective
ADV adverb
CONJ conjunction
DET determiner
FEM feminine
IND indicative
INF infinitive
GER gerund
LVC light verb construction
MASC masculine
MWE multiword expression
.45lQ N noun
NLP natural language processing
PAST past tense
PL plural
POS part of speech
PREP preposition
PRES present tense
PRON pronoun
SG singular
SUBJ subjunctive
V verb

Appendix

List of abbreviations used in the appendix

.45>lQ 1/2/3 pers 1st/2nd/3rd person
adj adjective
adv adverb
conj conjunction
det determiner
fem feminine
ger gerund
ind indicative
inf infinitive
masc masculine
n noun
.45>lQ past past tense
pl plural
pos possessive
pp past participle
pres present tense
prep preposition
refl v reflexive verb
pron pronoun
sg singular
subj subjunctive
v verb

The following three appendices present the Spanish data set used in this article classified according to our taxonomy. It shall be noted that the translations of s not always result in s in the target language, nor in the same syntactic class.

Appendix A: Spanish Fixed MWEs data set

Table 1: Adjectival phrases.

	Spanish MWE	**PoS pattern in Spanish**	**English translation**
1	*a cuadros*	prep + n	plaid
2	*a rayas*	prep + n	striped
3	*como puños*	adv + n	like daggers
4	*como una catedral*	adv + det + n	huge
5	*contante y sonante*	adj + conj + adj	hard cash
6	*corriente y moliente*	adj + conj + adj	plain ordinary
7	*de gala*	prep + n	gala
8	*de pared*	prep + n	wall
9	*de segunda mano*	prep + adj + n	second hand
10	*en directo*	prep + n	live
11	*en falso*	prep + adj	lie
12	*en jarras*	prep + n	on hips
13	*en vivo*	prep + adj	live
14	*mondo y lirondo*	adj + conj + adj	plain and simple

Table 2: Adverbial expressions.

c@ >p2cmQ@p3cm

Spanish MWE PoS pattern in Spanish English translation
1 *a bote pronto* prep + n (masc; sg) + adj (masc; sg) out of the blue
2 *a caballo* prep + n (masc; sg) on horseback
3 *a escondidas* prep + pp (fem; pl) behind somebody's back
4 *a fondo* prep + n (masc; sg) in depth
5 *a grito pelado* prep + n (masc; sg) + adj (masc; sg) at the top of one's lungs
6 *a gusto* prep + n (masc; sg) at ease
7 *a la carrera* prep + det (fem; sg) + n (fem; sg) in a rush
8 *a la fuerza* prep + det (fem; sg) + n (fem; sg) by force
9 *a la perfección* prep + det (fem; sg) + n (fem; sg) to perfection
10 *a la vez* prep + det (fem; sg) + n (fem; sg) all at once
11 *a la vista* prep + det (fem; sg) + n (fem; sg) in sight
12 *a las mil maravillas* prep + det (fem; pl) + adj + n (fem; pl) perfectly
13 *a manos llenas* prep + n (fem; pl) + adj (fem; pl) hand over fist
14 *a medias* prep + adj (fem; pl) halfway
15 *a oscuras* prep + adj (fem; pl) in the dark
16 *a secas* prep + adj (fem; pl) plainly
17 *a tientas* prep + n (fem; pl) blindly
18 *a toda velocidad* prep + adj (fem; sg) + n (fem; sg) at full speed
19 *al por mayor* prep + det (masc; sg) + prep + adj (masc; sg) wholesale
20 *codo con codo* n (masc; sg) + prep + n (masc; sg) elbow-to-elbow
21 *con las manos en la masa* prep + det (fem; pl) + n (fem; pl) + prep + det (fem; sg) + n (fem; sg) red-handed
22 *contra reloj* prep + n (masc; sg) against the clock
23 *con una mano delante y otra detrás* prep + det (fem; sg) + n (fem; sg) + adv + conj + adj (fem; sg) + adv from hand to mouth
24 *de buenas* prep + adj (fem; pl) with all one's heart
25 *de cabo a rabo* prep + n (masc; sg) + prep + n (masc; sg) head to tail
26 *de golpe y porrazo* prep + n (masc; sg) + conj + n (masc; sg) all of a sudden
27 *de reojo* prep + n (masc; sg) out of the corner of one's eye

c@ >p2cml@Q		
28	*en breve*	prep + adj (masc; sg) shortly/in due course
29	*en consecuencia*	prep + n (fem; sg) consequently
30	*en definitiva*	prep + adj (fem; sg) in conclusion
31	*en el acto*	prep + det (masc; sg) + n (masc; sg) in the act
32	*en líneas generales*	prep + n (fem; pl) + adj (fem; pl) by and large
33	*en pocas palabras*	prep + adj (fem; pl) + n (fem; pl) in a nutshell
34	*en secreto*	prep + n (masc; sg) in secret
35	*en suma*	prep + n (fem; sg) in short
36	*en un santiamén*	prep + det (masc; sg) + n (masc; sg) in a flash
37	*más o menos*	adv + conj + adv more or less
38	*ni más ni menos*	conj + adv + conj + adv
39	*para colmo*	prep + n (masc; sg) to top it all
40	*por casualidad*	prep + n (fem; sg) by chance
41	*por cierto*	prep + adj (masc; sg) by the way
42	*por consiguiente*	prep + adj (masc; sg) hence
43	*por descontado*	prep + pp (masc; sg) needless to say
44	*por el contrario*	prep + det (masc; sg) + adj (masc; sg) on the contrary
45	*por supuesto*	prep + adj (masc; sg) of course
46	*sin embargo*	prep + n (masc; sg) nevertheless
47	*sin más ni más*	prep + adv + conj + adv just like that
48	*sin ton ni son*	prep + n (masc; sg) + adv + n (masc; sg) without rhyme or reason
49	*una barbaridad*	det (fem; sg) + n (fem; sg) quite a lot

Table 3: Conjunctional phrases.

	Spanish MWE	**PoS pattern in Spanish**	**English translation**
1	*a fin de que*	prep + n (masc; sg) + prep + conj	in order to
2	*a medida que*	prep + n (fem; sg) + conj	as
3	*a menos que*	prep + adv + conj	unless
4	*así que*	adv + conj	consequently
5	*con tal de que*	prep + adv + prep + conj	as long as
6	*mientras que*	adv + conj	while
7	*siempre que*	adv + conj	whenever
8	*tan pronto como*	adv + adv + conj	as soon as
9	*visto que*	adj + conj	since
10	*ya que*	adv + conj	because

Table 4: Complex nominals.

c@ >p3cmp5cm@ Q

Spanish MWE	PoS pattern	in Spanish English translation
1 *abrebotellas*	n (masc; sg/pl)	bottle opener
2 *aguafiestas*	n (masc/fem; sg/pl)	spoilsport
3 *cascanueces*	n (masc; sg/pl)	nutcracker
4 *correveidile*	n (fem/masc; sg)	tell-tale
5 *lavavajillas*	n (masc; sg/pl)	diswasher
6 *limpiacristales*	n (fem/masc; sg/pl)	window cleaner
7 *rascacielos*	n (masc; sg/pl)	skyscrapper
8 *sacacorchos*	n (masc; sg/pl)	bottle opener
9 *soplagaitas*	n (fem/masc; sg/pl)	dumbbell
10 *pinchadiscos*	n (masc/fem; sg/pl)	disc jockey
11 *complejo de Edipo*	n (masc; sg) + prep + n (masc; sg)	Oedipus complex
12 *el día a día*	det (masc; sg) n (masc; sg) + prep + n (masc; sg)	everyday life
13 *el día del juicio final*	det (masc; sg) + n (masc; sg) + prep + det (masc; sg) + n (masc; sg) + adj (masc; sg)	doomsday
14 *gripe aviar*	n (fem; sg) + adj (fem; sg)	avian influenza
15 *la flor y la nata*	det (fem; sg) + n (fem; sg) + conj + det (fem; sg) + n (fem; sg)	cream of the crop
16 *la gran pantalla*	art (fem; sg) + adj (fem; sg) + n (fem; sg)	the big screen
17 *la teoría de la relatividad*	det (fem; sg) + n (fem; sg) + prep + det (fem; sg) + n (fem; sg)	theory of relativity
18 *mucho ruido y pocas nueces*	adj (masc; sg) + n (masc; sg) + conj + adj (fem; pl) + n (fem; pl)	much ado about nothing
19 *perro ladrador, poco mordedor*	n (masc; sg) + adj (masc; sg) + adv + adj (masc; sg)	his bark is worse than his bite
20 *sentido del ridículo*	n (masc; sg) + prep + n (masc; sg)	self-concious
21 *síndrome de down*	n (masc; sg) + prep + n (masc; sg)	Down Syndrome
22 *vergüenza ajena*	n (fem; sg) + adj (fem; sg)	feel embarrassment for
23 *síndrome de Estocolmo*	n (masc; sg) + prep + n (masc; sg)	Stockholm Syndrome

Table 5: Proper names.

c>p3cmQp3cm

Spanish MWE PoS pattern in Spanish English translation
1 *Air Jordan* n (masc; sg) + n (masc; sg) Air Jordan
2 *Al Capone* n (masc; sg) + n (masc; sg) Al Capone
4 *América Latina* n (fem; sg) + adj (fem; sg) Latin America
5 *Amnistía Internacional* n (fem; sg) + adj (fem; sg) Amnesty International
6 *Banco Central Europeo* n (masc; sg) + adj (masc; sg) + adj (masc; sg) European Central Bank
7 *Billy el Niño* n (masc; sg) + det (masc; sg) + n (masc; sg) Billy the Kid
8 *Buenos Aires* adj (masc; pl) + n (masc; pl) Buenos Aires
9 *Costa Rica* n (fem; sg) + adj (fem; sg) Costa Rica
10 *Cruz Roja* n (fem; sg) + adj (fem; sg) Red Cross
11 *el Cordobés* det (masc; sg) + adj (masc; sg) el Cordobés
12 *El Greco* det (masc; sg) + adj (masc; sg) El Greco
13 *El Pelusa* det (masc; sg) + n (fem; sg) el Pelusa
14 *El Principito* det (masc; sg) + n (masc; sg) The Little Prince
15 *Gran Bretaña* adj (fem; sg) + n (fem; sg) Great Britain
15 *José María* n (masc; sg) + n (fem; sg)
16 *La Paz* det (fem; sg) + noun (fem; sg) La Paz
17 *La sombra del viento* det (fem; sg) + n (fem; sg) + prep + det (masc; sg) + n (masc; sg) The Shadow of the Wind
18 *Lawrence de Arabia* n (masc; sg) + prep + n (fem; sg) Lawrence of Arabia
19 *Lord Byron* n (masc; sg) + n (masc; sg) Lord Byron
20 *Los Ángeles* det (masc; pl) + n (masc; pl) Los Angeles
21 *Manchester United* n + adj Manchester United
22 *María José* n (fem; sg) + n (masc; sg)
23 *Médicos Sin Fronteras* n (masc; pl) + prep + n (fem; pl) Doctors Without Borders
24 *Mona Lisa* n (fem; sg) + n (fem; sg) Mona Lisa
25 *Nueva York* adj (fem; sg) + n (fem; sg) New York
26 *Nueva Zelanda* adj (fem; sg) + n (fem; sg) New Zealand
27 *Osa Mayor* n (fem; sg) + adj (fem; sg) Ursa Major
28 *Países Bajos* n (fem; sg) + adj (fem; sg) the Netherlands
29 *Papá Noel* n (masc; sg) + n (masc; sg) Father Christmas
30 *Real Academia Española* adj (fem; sg) + n (fem; sg) + adj (fem; sg)
31 *Real Madrid* adj (masc; sg) + n (masc; sg) Real Madrid
32 *Reino Unido* n (masc; sg) + adj (masc; sg) United Kingdom
33 *República Dominicana* n (fem; sg) + adj (fem; sg) Dominican Republic
34 *San Salvador* adj (fem; sg) + n (masc; sg) San Salvador
35 *Unión Europea* n (fem; sg) + adj (fem; sg) European Union

Table 6: Prepositional phrases.

	Spanish MWE	PoS pattern in Spanish	English translation
1	*por culpa de*	prep + n (fem; sg) + prep	because of
2	*a pesar de*	prep + n (masc; sg) + prep	in spite of
3	*al margen de*	prep + det (masc; sg) + n (masc; sg) + prep	apart from
4	*con miras a*	prep + n (fem; sg) + prep	looking to
5	*de conformidad con*	prep + n + prep	according to
6	*en contra de*	prep + n (fem; sg) + prep	in opposition to
7	*en cuanto a*	prep + adverb + prep	with regard to
8	*en detrimento de*	prep + n (masc; sg) + prep	at the expense of
9	*en relación con*	prep + n + prep	in relation to
10	*respecto a*	n + prep	in relation to

Table 7: Sentential expressions.

cp3cmQp3cm

	Spanish MWE	PoS pattern in Spanish	English translation
1	*cuando el río suena, agua lleva*	conj + det (masc; sg) + n (masc; sg) + v (3rd pers; sg) + n (fem; sg) + v (3rd pers; sg)	where there's smoke, there's fire
2	*cuando las ranas críen pelo*	adv + det (fem; pl) + n (fem; pl) + v (3rd pers; pl) + n (masc; sg)	when pigs fly
3	*dime con quién andas y te diré quién eres*	v (2nd pers; sg) + prep + pron + v (2nd pers; sg) + conj + pron + v (1st pers; sg) + pron + v (2ª pers; sg)	birds of a feather flock together
4	*más vale tarde que nunca*	adv + v (3rd pers; sg) + adv + conj + adv	better late than never

Appendix B: Spanish Semi-fixed MWEs data set

Table 8: Adjectival compounds.

	Spanish MWE	PoS pattern in Spanish	English translation
1	*agridulce*	adj (masc/fem; sg)	sweet-and-sour/bittersweet
2	*boquiabierto*	adj (masc; sg)	open-mouthed
3	*cabizbajo*	adj (masc; sg)	downcast
4	*cejijunto*	adj (masc; sg)	unibrow
5	*drogadicto*	adj (masc; sg)	drug addict
6	*hispanohablante*	adj (masc/fem; sg)	Spanish-speaking
7	*narcotraficante*	adj (masc/fem; sg)	drud dealer/drug trafficker
8	*patidifuso*	adj (masc; sg)	astonished
9	*pelirrojo*	adj (masc; sg)	redheaded
10	*vasodilatador*	adj (masc; sg)	vasodilator

Table 9: Adjectival phrases.

lXll

Spanish MWE	PoS pattern in Spanish	English translation
1 *de primera mano*	prep + adj + n	first hand
2 *sano y salvo*	adj + conj + adj	safe and sound

Table 10: Adverbial expressions.

lXll

Spanish MWE	PoS pattern in Spanish	English translation
1 *a golpes*	prep + n (masc; pl)	violently

Table 11: Complex nominals.

c>p2.5cmQp3cm

	Spanish MWE PoS pattern in Spanish English translation
1	*la ley de la jungla* det (fem; sg) + n (fem; sg) + prep + det (fem; sg) + n (fem; sg) law of the jungle
2	*anillo de compromiso* n (masc; sg) + prep + n (masc; sg) engagement ring
3	*bicicleta estática* n (fem; sg) + adj (fem; sg) exercise bike
4	*bocacalle* n (fem;sg) side-street
5	*bomba nuclear* n (fem; sg) + adj (fem; sg) nuclear bomb
6	*café con leche* n (masc; sg) + prep + n (fem; sg) coffee with milk
7	*campo de concentración* n (masc; sg) + prep + n (fem; sg) concentration camp
8	*centro de salud* n (masc; sg) + prep + n (fem; sg) health center
9	*cinta de correr* n (fem; sg) + prep + inf treadmill
10	*ciudad dormitorio* n (fem; sg) + n (masc; sg) dormitory town
11	*complejo de inferioridad* n (masc; sg) + prep + n (fem; sg) inferiority complex
12	*crema de manos* n (fem; sg) + prep + n (fem; pl) hand cream
13	*cuenta de débito* n (fem; sg) + prep + n (masc; sg) debit account
14	*cuenta de resultados* n (fem; sg) + prep + n (masc; pl) profit and loss account
15	*cuento chino* n (masc; sg) + adj (masc; sg) a tall tale
16	*deporte de aventura* n (masc; sg) + prep + n (fem; sg) adventure sport
17	*diente de león* n (masc; sg) + prep + n (masc; sg) dandelion
18	*disco pirata* n (masc; sg) + n (masc; sg) pirate CD
19	*fin de semana* n (masc; sg) + prep + n (fem; sg) weekend
20	*goma de borrar* n (fem; sg) + prep + inf eraser
21	*guerra civil* n (fem; sg) + adj (fem; sg) civil war
22	*hombre lobo* n (masc; sg) + n (masc; sg) werewolf
23	*hueso duro de roer* det (masc; sg) + n (masc; sg) + adj (masc; sg) + prep + inf hard nut to crack
24	*impuesto revolucionario* n (masc; sg) + adj (masc; sg) revolutionary tax
25	*infarto de miocardio* n (masc; sg) + prep + n (masc; pl) myocardial infarction

c>p3cmQp3cm
26 *la gallina de los huevos de oro* det (fem; sg) + n (fem; sg) + prep + det (masc; pl) + n (masc; pl) + prep + n (masc; sg) cash cow
27 *la ley del más fuerte* det (fem; sg) + n (fem; sg) + prep + det (masc; sg) + adv + adj (masc; sg) survival of the fittest
28 *lobo con piel de cordero* noun (masc; sg) + prep + noun (fem; sg) + prep + noun (masc; sg) wolf in sheep's clothing
29 *mesa camilla* n (fem; sg) + n (fem; sg) round table
30 *momento clave* n (masc; sg) + n (fem; sg) key moment
31 *niño mimado* n (masc; sg) + adj (masc; sg) blue-eyed boy
32 *niño prodigio* n (masc; sg) + n (masc; sg) child prodigy
33 *patata caliente* noun (fem; sg) + adj (fem; sg) hot potato
34 *perro de caza* n (masc; sg) + prep + n (fem; sg) hunting dog
35 *raíz cuadrada* n (fem; sg) + adj (fem; sg) square root
36 *realidad virtual* n (fem; sg) + adj (fem; sg) virtual reality
37 *renta per cápita* n (fem; sg) + prep + n (fem; sg) income per capita
38 *ruleta rusa* n (fem; sg) + adj (fem; sg) Russian roulette
39 *salto mortal* n (masc; sg) + adj (masc; sg) somersault
40 *sentimiento de culpa* n (masc; sg) + prep + n (fem; sg) guilt
41 *tarjeta de crédito* n (fem; sg) + prep + n (masc; sg) credit card
42 *tortilla de patata* n (fem; sg) + prep + n (fem; sg) Spanish omelette
43 *zumo de naranja* n (masc; sg) + prep + n (fem; sg) orange juice

Appendix C: Spanish Flexible MWEs data set

Table 12: Verbal phrases.

c>p2.5cmQp3.5cm

Spanish MWE PoS pattern in Spanish English translation
1 *coger el toro por los cuernos* v + det (masc; sg) + n (masc; sg) + prep + det (masc; pl) + n (masc; pl) to take the bull by the horns
2 *echar por tierra* v + prep + n (fem; sg) to upset the applecart
3 *empezar la casa por el tejado* v + det (fem; sg) + n (fem; sg) + prep + det (masc; sg) + n (masc; sg) to put the cart before the horse
4 *estar como unas castañuelas* v + adv + det (fem; pl) + n (fem; pl) to be tickled pink
5 *ir de guatemala a guatepeor* v + prep + n (fem; sg) + prep + n (masc; sg) out of the frying pan and into the fire
6 *ni pinchar ni cortar* conj + v + conj + v to cut no ice
7 *ser de armas tomar* v + prep + n (fem; pl) + verb to be someone to be reckoned with
8 *ser el ojito derecho* v + det (masc; sg) + n (masc; sg) + adj (masc; sg) to be the apple of one's eye
9 *ser harina de otro costal* v + n (fem; sg) + prep + adj (masc; sg) + n (masc; sg) to be a horse of a different colour
10 *ser la crème de la crème* det (fem; sg) + n (fem; sg) + prep + det (fem; sg) + n (fem; sg) to be crème de la crème
11 *vivir a cuerpo de rey* v + prep + n (masc; sg) + prep + n (masc; sg) to live high on the hog

Table 13: Sentential expressions.

c>p2.5cmQp3cm

Spanish MWE PoS pattern in Spanish English translation
1 *la gota que colma el vaso* det (fem; sg) + n (fem; sg) + conj + v (3rd pers; sg) + det (masc; sg) + n (masc; sg) straw that breaks the camel's back

Table 14: Adjectival phrases.

cXll

Spanish MWE PoS pattern in Spanish English translation
1 *de cuidado* prep + n (masc; sg) dangerous
2 *de ensueño* prep + n (masc; sg) fantastic

Table 15: Adjectives with a governed prepositional phrase.

cXll

Spanish MWE PoS pattern in Spanish English translation
1 *adicto a* adj (masc; sg) + prep addicted to
2 *aficionado a* adj (masc; sg) + prep fond of
3 *apto para* adj (masc; sg) + prep suitable for
4 *aspirante a* adj (masc/fem; sg) + prep candidate for
5 *carente de* adj (masc/fem; sg) + prep deprived of
6 *casado con* adj (masc; sg) + prep married to/with
7 *celoso de* adj (masc; sg) + prep jealous of
8 *culpable de* adj (masc/fem; sg) + prep guilty of
9 *dependiente de* adj (masc/fem; sg) + prep dependent on
10 *exento de* adj (masc; sg) + prep exempt from
11 *interesado en* adj (masc; sg) + prep interested in
12 *preocupado por* adj (masc; sg) + prep worried about
13 *sospechoso de* adj (masc; sg) + prep suspected of

Table 16: Adverbial expression.

c>Q@lp3.3cm

Spanish MWE PoS pattern in Spanish English translation
1 *a mi/tu/su/nuestro/vuestro entender* prep + pos + n (masc; sg) by my/your/her/his/our/ their understanding

Table 17: Nouns with a governed prepositional phrase.

c>Qlp3.3cm

	Spanish MWE PoS pattern	in Spanish English translation
1	*actitud con/hacia/respecto de*	noun (fem; sg)+ prep attitude with/towards/regarding
2	*amenaza de*	n (fem; sg) + prep threat of
3	*asalto a/de*	n (masc; sg) + prep assault to/on
4	*confianza en*	n (fem; sg) + prep trust in
5	*esperanza de*	n (fem; sg) + prep hope to
6	*interés por*	n (masc; sg) + prep interest in
7	*olor a*	n (masc; sg) + prep smell of
8	*prohibición de*	n (fem; sg) + prep prohibition of
9	*sabor a*	n (masc; sg) + prep taste of
10	*salida de*	n (fem; sg) + prep exit of
11	*traducción a*	n (fem; sg) + prep translation to
12	*veto a*	n (fem; sg) + prep ban on

Table 18: Light verb constructions.

c>Qlp3cm

	Spanish MWE PoS pattern	in Spanish English translation
1	*cantar las cuarenta*	v + det (fem; pl) + adj (fem pl) to haul over the coals
2	*comer la olla*	v + det (fem; sg) + n (fem; sg) to talk someone into something
3	*cortar el bacalao*	v + det (masc; sg) + n (masc; sg) to be the big cheese/big fish
4	*dar acidez*	v + n (fem; sg) to produce heartburn
5	*dar ánimos*	v + n (masc; pl) to cheer up
6	*dar calor*	v + n (masc; sg) to keep warm
7	*dar carpetazo*	v + n (masc; sg) to put an end to
8	*dar esquinazo*	v + n (masc; sg) to give the slip
9	*dar la palabra*	v + det (fem; sg) + n (fem; sg) to give the floor to
10	*dar la tabarra*	v + det (fem; sg) + n (fem; sg) to pester
11	*dar plantón*	v + n (masc; sg) to stand [sb] up
12	*dar suerte*	v + n (fem; sg) to give [sb] luck

c>Qlp3cm

13 *dar un beso* v + det (masc; sg) + n (masc; sg) to give a kiss
14 *dar una patada* v + det (fem; sg) + n (fem; sg) to kick
15 *dar un paseo* v + det (masc; sg) + n (masc; sg) to go for a walk
16 *dar un puñetazo* v + det (masc; sg) + n (masc; sg) to punch
17 *despertar el apetito* v + det (masc; sg) + n (masc; sg) to awaken one's apettite
18 *echar la siesta* v + det (fem; sg) + n (fem; sg) take a nap
19 *echar un cable* v + det (masc; sg) + n (masc; sg) give a hand
20 *empinar el codo* v + det (masc; sg) + n (masc; sg) to bend one's elbow
21 *hacer alusión* v + n (fem; sg) to make an allusion
22 *hacer añicos* v + n (masc; pl) to break into pieces
23 *hacer gracia* v + n (fem; sg) to be funny
24 *hacer ilusión* v + n (fem; sg) to look forward to
25 *hacer la compra* v + det (fem; sg) + n (fem; sg) to do the shopping
26 *hacer la pelota* v + det (fem;sg) + n (fem; sg) to suck up to
27 *hacer un trato* v + det (masc; sg) + n (masc; sg) to make a deal
28 *hacer una foto* v + det (fem; sg) + n (fem; sg) to take a picture
29 *hacer una oferta* v + det (fem; sg) + n (fem; sg) to make an offer
30 *hacerse ilusiones* refl v + n (fem; pl) to get one's hopes up
31 *levar anclas* v + n (fem; pl) to weigh anchor
32 *llamar la atención* v + det (fem; sg) + n (fem; sg) to attract one's attention
33 *pasar la pelota* v + det (fem; sg) + n (fem; sg) to pass the buck
34 *ponerse las pilas* refl v + det (fem; pl) + n (fem; pl) to get one's act together
35 *sacar pecho* v + n (masc; sg) to stick your chest out
36 *sufrir las consecuencias* v + det (fem; pl) + n (fem; pl) to suffer the consequences
37 *tener gana* v + n (fem; sg) to be hungry
38 *tener ganas* v + n (fem; pl) to feel like
39 *tomar el pelo* v + det (masc; sg) + n (masc; sg) to tease [someone]
40 *tomar el sol* v + det (masc; sg) + n (masc; sg) to sunbathe
41 *tomar partido* v + n (masc; sg) to take sides
42 *tomar una decisión* n + det (fem; sg) + n (fem; sg) to make a decision

Table 19: Periphrastic constructions. Periphrastic constructions do not have straightforward English translations. The ones give here are an indication of what the usually mean but the translations will depend on the verb appearing in a non-finite form in the periphrasis.

	Spanish MWE	**PoS pattern in Spanish**	**English translation**
1	*acabar de + inf*	v + prep + inf	to finish to
2	*andar + ger*	v + ger	to be doing
3	*deber + inf*	v + inf	to have to
4	*deber de + inf*	v + prep + inf	to may have
5	*empezar a + inf*	v + prep + inf	to begin to
6	*estar por + inf*	v + prep + inf	to be about to
7	*haber de + inf*	v + prep + inf	to have to
8	*haber que + inf*	v + pron + inf	to have to
9	*ir + ger*	v + ger	to begin/be doing
10	*ir a + inf*	v + prep + inf	to go to
11	*llegar a + inf*	v + prep + inf	to manage to
12	*llevar + ger*	v + ger	to have been doing
13	*llevar + pp*	v + pp	to have done
14	*poder + inf*	v + inf	to be able to
13	*sacar a + inf*	v + prep + inf	to take someone out to
15	*seguir + ger*	v + ger	to continue doing
16	*tener que + inf*	v + pron + inf	to have to
17	*venir + ger*	v + ger	to have been doing
18	*venir a + inf*	v + prep + inf	to be
19	*volver a + inf*	v + prep + inf	to do something again

Table 20: Verbal phrases.

c>p2.5cmQp3cm

Spanish MWE	**PoS pattern in Spanish**	**English translation**
1 *dar por sentado*	v + prep + adj (masc; sg)	take for granted
2 *entrar al trapo*	v + prep + det (masc; sg) + n (masc; sg)	to respond to provocations
3 *estar al pie del cañón*	v + prep + det (masc; sg) + n (masc; sg) + prep + det (art; sg) + n (masc; sg)	to be ready and waiting
4 *estar en Babia*	v + prep + n (fem; sg)	to be daydreaming
5 *estar en las nubes*	v + prep + det (fem; pl) + noun (fem; pl)	to be in the clouds
6 *hacer una montaña de*	v + det (fem; sg) + n (fem; sg) + prep + det (masc; sg) + n (masc; sg) + prep + n (fem; sg)	make a mountain out of molehill
7 *irse de la lengua*	refl v + prep + det (fem; sg) + n (fem; sg)	to let the cat out of the bag
8 *irse de rositas*	refl v + prep + n (fem; pl)	to get off scot free
9 *irse por las ramas*	refl v + prep + det (fem; pl) + n (fem; pl)	to beat around the bush
10 *llamar a la puerta equivocada*	v + prep + det (fem; sg) + n (fem; sg) + adj (fem; sg)	to bark up the wrong tree
11 *salir al paso*	v + prep + det (masc; sg) + n (masc; sg)	to refute
12 *salir de cuentas*	v + prep + n (fem; pl)	to be due
13 *salir de marcha*	v + prep + n (fem; sg)	go partying
14 *saltar a la comba*	v + prep + det (fem; sg) + n (fem; sg)	to skip rope
15 *ser fiel a*	v + adj (masc/fem; sg) + prep	to be loyal to

Table 21: Verbs with a governed prepositional phrase.

	Spanish MWE	PoS pattern	English translation
1	*abstenerse de*	refl v + prep	to refrain yourself from
2	*acordarse de*	refl v + prep	to remember
3	*amenazar con*	v + prep	to threaten to
4	*arrepentirse de*	refl v + prep	to regret
5	*atenerse a*	refl v + prep	to stick to
6	*confiar en*	v + prep	to trust in
7	*contribuir a*	v + prep	to contribute to
8	*creer en*	v + prep	to believe in
9	*cuidar de*	v + prep	to take care of
10	*empeñarse en*	refl v + prep	to insist on
11	*engancharse a*	refl v + prep	to get hooked on
12	*entender de*	v + prep	to know about
13	*eximir de*	v + prep	to exempt from
14	*gozar de*	v + prep	to enjoy
15	*hablar de/sobre/acerca de*	v + prep	to talk about/of
16	*interesarse por*	refl v + prep	to be interested in
17	*oler a*	v + prep	to smell like
18	*pelear por*	v + prep	to fight for
19	*quedarse con*	refl v + prep	to keep
20	*referirse a*	refl v + prep	to refer to
21	*viajar a/hacia/hasta*	v + prep	to travel to/towards