Université François-Rabelais Tours

Information Technologies for Business Intelligence Master Programme

Business Intelligence Seminar

Representation and Identification of Multiword Expressions in different Spanish Dialects

Authors:
Alejandro Arauco
Diana Bogantes
Alejandro Rodríguez
Eric Rodríguez

Supervisor:
Agata Savary
Laboratoire d'Informatique
3 place Jean-Jaurès
41029 Blois – FRANCE
Tél.: 02.47.36.14.14
Fax: 02.47.36.14.36
agata.savary@univ-tours.fr

29 June 2015
Technical Report n°314, 38 pages
Extended version of a poster to be presented at the 5th meeting of the IC1207 COST action PARSEME in Iasi, Romania

"Hoy es día de tianguis. Voy a ir temprano, no sea que (1) *me coman el mandado* y me quede sin lugar para mi puesto. La verdad, es que me ha (2) *ido de pelos* con las ventas, pero eso sí, las (3) *mordidas* están (4) *al por mayor* con los inspectores, porque luego quieren (5) *hacerla de tos* con los permisos y demás; no falta el que dice que no tienes bien tus papeles y que hay que (6) *mocharse con una lana*. ¡Qué barbaridad! Así no se puede trabajar bien, (7) *a gusto*, con esos (8) *coyotes* que (9) *lo traen a uno asoleado*, eso que se dicen tus (10) *cuates*, ¡qué tal si no lo fueran! ¡Ya me imagino!"

Nancy Altamirano Ceballos

Contents

1 Introduction

1.1 Multiword Expressions

Muti-word expressions (MWE) encompass a wide range of linguistically related phenomena that share the criterion of being composed of two or more words, either adjacent or separate [Attia, 2006]. All these kinds of expressions share the fact that the semantic meaning cannot be deducted from their components. A typical example in English of this type of expressions is '*to spill the beans*'. Its meaning, to reveal a secret, has no (or limited) correlation to '*spill*' and '*beans*'. These kinds of expressions are very common among languages and dialects, however they can vary from one to another either in meaning or in the actual expression used to express a particular sentiment.

For example, in Spanish from Costa Rica [CR][1] the expression '*estar limpio*' (#9)[2] which literally translates to '*be clean*', can also mean '*to be out of money*'. However, this expression doesn't have that meaning in other dialects like Spanish from Colombia [COL] or Mexico [MEX]. It can also be the case that there is a similar MWE expression with the same meaning in another dialect. For example, '*estar aguja*' [PE] that translates to '*be needle*' also means '*to be out of money*'. Another scenario we can find is that the same expression has the same meaning in several dialects, e.g. '*echar los perros*' (#31), '*throw the dogs*', which means '*to flirt*' in Colombia, Costa Rica and Mexico. And the last case is when the same MWE has different meanings in all dialects, e.g. '*ponerse las pilas*' (#4) can mean '*o start doing something seriously* [COL], '*to do things in a better way*' [CR], '*to be more active*' [MEX] or '*to do things faster*' [PE].

Handling this kind of expressions has been and is still a challenge for applications in natural language processing (NLP). It's also well known that in morphologically rich languages, like Spanish, the challenge is bigger. Spanish, compared to English, has extensive possibilities of grammatical inflections, especially in the conjugation of verbs. Looking at the classification of the MWE in [Attia, 2006], it is possible to get an idea of the reason why the morphological richness impacts the way we handle MWE.

In this classification model, the first type of MWEs are the **fixed expressions**. They can be seen as single word that happen to have a space [Attia, 2006] like *Costa Rica*. This expressions are morphologically rigid, meaning that they omit inflections. The next group are the **semi-fixed expressions** that are affected by morphological transformations. For example the compound noun

[1] When a Spanish example is given thought this document, the dialect of origin is shown between square brackets to specify that the MWE or the meaning are valid only in that dialect. If nothing is specified then it applies for all four dialects: Colombia, Costa Rica, Mexico and Perú

[2] With every example we include the number of that phrase in the Appendix section for further morphologic analysis details

'*kick the bucket*', which means '*to die*' can be found as '*kick/kicks/kicked*' the bucket. Hence, this kinds of MWEs are more complex to handle by NLP applications, since having the list of expressions and doing an exact match is not enough to identify them, as it is for the first group. The complexity increases if the language is morphologically rich, because it enables more variations of the MWE. For example, '*metió la pata*' (#34) that means '*to screw up*' [COL, CR] (but its word by word translation is 'insert the leg') can be transformed to '*meterá/metió/ hubo metido/metería/ mete/ meter/ha metido/ habré metido/ mete/metiera/ meterás/ meterá/había metido/ haya metido (etc.) la pata*'. We get this variations by only taking into account the available tenses of the verbs in Spanish: present, indicative, imperfect, preterit, future, conditional, imperative, present subjunctive, imperfect subjunctive, gerund and past participle.

The third group are the **flexible expressions**, MWEs that can be reordered or that accept other words or expressions to appear between the components. For example, '*Hablar paja*' (#20), '*talk straw*' that means '*small talk* [COL], can be transformed to '*habla pura/solo paja*' where '*pura* and *solo*' mean '*only*'. The fourth group identified by [Sag et al., 2002] are the **institutionalized phrases** that are described in Table 1. These are MWEs semantically and syntactically compositional, meaning that there is a clear relation between the meaning of the components and the meaning of the MWE, but have been widely used compared to other expressions, converting them into parts of the language. E.g '*salt and pepper*'.

1.2 Related Work

The interest in the challenges that MWE bring to NLP applications continues to grow within the NLP community and there are different topics that are currently being studied. For example:

- **Linguistic analysis of MWEs**. This field deals with the definition of appropriate linguistic descriptions for these expressions. Additional work has been done in this area related to the creation of processors that output the morphological analysis of a certain class of MWE received as input [Oflazer et al., 2004].
- **Lexical resources and ontologies to express MWEs**. The main tasks of this line of work is defining strategies for encoding MWEs in lexical resources like electronic dictionaries of MWEs, table-like structures to store the properties of MWEs and schemas for representing the different combinations of syntactic and lexical variations that can occur in a MWE. In [Al-Haj, 2009] there is a summary of some of the works done on this area and the proposal of a new schema for MWE representation.
- **Automatic identification and extraction of MWEs**. This task is

done on written corpora, and there have been many research and implementation approaches in different languages as described in [Al-Haj, 2009], [de Caseli et al., 2010], [Attia et al., 2010] and [Ramisch, 2015].

1.3 Research Goals

In this work we take the schema proposed in [Itai and Wintner, 2008] to define our own model to represent Spanish MWEs when different dialects are considered. This entails extending the model in a way that is possible to: (1) represent the dialects in which the MWE is valid and with which meaning (considering that the same expression can mean different things along the dialects), (2) link MWEs that are different in form but same in meaning in one or more Spanish dialects, (3) specify whether or not the MWE has agreement restrictions on, for example, gender and number between words of the MWE, and (4) stipulate if the MWE allows inflections, substitutions or additions, in which level and in which elements within the MWE these changes are allowed.

Moreover, we describe the process used for the creation of a corpus for these expressions. For this web crawling and crowd sourcing techniques were joined in order to start building a corpus that can then be used in future related works. The schema for the MWE representation allows the possibility of including references to the corpus document where a specific MWE expression can be found. Once the schema and the corpus process were defined we used them to create two main resources: (1) an XML document with the complete analysis of a set of Spanish MWE and (2) a corpus containing the web pages retrieved by the process where MWE can be found .

The remainder of this document is structured as follows: in Section 2 we discuss about the properties, challenges and dialect-oriented issues of MWEs as well as the motivation for handling them in NLP applications. In Section 3 we describe some properties of Spanish MWE, then in Section 4 we introduce the proposed schema to represent those expressions. In section 5 the details of the corpus creation process are described, while in section 6 we show an example of extraction from the created corpus. In 7 section we show some statistics of the MWE database that we created and in section 8 we evaluate and describe the limits of the XML schema and the corpus. Finally the conclusions and future work are presented in Section 9.

2 Challenges and dialect-oriented issues of MWEs

As was briefly mentioned before, the handling and treatment of MWEs is an open and challenging problem. Nowadays the interest continues to grow and an important number of researchers are providing solutions in different ways. Determining and classifying what is considered a MWE inside text corpora is a challenging task that requires specialised knowledge and skills in the language

of the text. This task is commonly performed by humans, nonetheless, one of the areas where NLP applications have been developed is the automatic identification and extraction of these expressions when present in a text.

2.1 How humans can detect MWEs

In order to design an automatic process for MWE recognition and extraction, first we need to understand how humans identify them. According to [Ramisch, 2015], one possible way to recognise MWEs is by applying simple linguistic tests, such as replacing one word in the expression for a synonym and reviewing the result. If the new phrase seems awkward to a native speaker of the language it means that we have found a MWE. For example, the Spanish phrase '*caer bien*' (#60) [CR, MEX] whose literal translation to English is to '*fall good*', means get on well / get along well. However, changing one word of the original phrase from '*caer bien*' to '*caer aceptable*' ('*fall acceptable*') for example, will drift the original meaning into an unnatural significance.

Another test for detecting MWEs mentioned in [Ramisch, 2015] is to perform a word to word translation of the phrase into another language. If the translation sounds weird, abnormal or even ungrammatical, the original expression is probably a MWE. For instance the expression '*pan de Dios*' in the sentence '*Lucía es un pan de Dios*' whose word-by-word translation is '*Lucía is a bread of God*' actually means '*Lucía is very kind*', therefore the literal translation lacks sense.

2.2 MWEs Properties

To have a better understanding of what is the main concern of the computational acquisition of MWEs it is important to study the main properties of these expressions. Table 1 shows a summary of the most important attributes of a MWE [Ramisch, 2015]

2.3 The big obstacle 'Computational treatment of MWEs'

As we have seen in previous sections, humans can smoothly identify a MWE by knowing its properties. An NLP system, however, doesn't know anything unless it is explicitly encoded. Therefore all information, like the assumptions about the general rules of the system, must be represented in a formal and explicit way. [Grégoire, 2009]

As a consequence, in [Grégoire, 2009] it's established that NLP systems, on the one hand, should contain –to adequately deal with large numbers of MWEs– (1) a suitable method for handling various types of MWEs in the grammar, and (2) a large number of lexical entries for MWEs compatible with the grammar. On the other hand, the lexical description of a MWE must have the same

Table 1: MWEs Properties

Nr.	Property	Meaning
1	Arbitrariness	Considered the most challenging property of a MWEs. It covers the case when a valid construction both syntactically and semantically is not acceptable simply because people do not talk that way.
2	Institutionalisation	Refers to the proportion of multiword expressions that are observed with higher frequency than any alternative lexicalization of the same concept
3	Limited semantic variability	In opposition to an ordinary word combination, MWEs do not undergo the same semantic compositionality rules. This characteristic is often expressed in terms of the following subproperties: **Non-compositionality**: the meaning of the whole expression often can't be directly inferred from the meaning of the parts composing it. **Non-substitutability**: is not possible to replace part of a MWE by a related synonym or equivalent word or construction. **No word-for-word translation** and **domain-specificity/idiomaticity**: a MWE is related to a specific sublanguage that might be a specialised scientific or technical domain.
4	Limited syntactic variability (non-modifiability)	Standard grammatical rules do not apply to some MWEs. This feature is demonstrated by its own sub-properties such as: **Extra-grammaticality** refers to an unpredictable and strange appearance of an expression for somebody who knows general syntactic rules. **Lexicalisation** expresses the fact that some words "belong together" in a single lexical unit.
5	Heterogeneity	As each MWE encompasses a large amount of distinct phenomena it's necessary to classify them by using one or more of the multiple methods available.

properties as a simple lexical item and furthermore: (1) a syntactic structure of the MWE, (2) a unique identification of the MWE components and (3) a listing of the MWE components in an order that is compatible with the syntactic structure.

As specified by [Gralinski et al., 2010], one of the main problems MWE related NLP systems is the need to deal with is the conflation of different surface realisations of the same underlying concept by the proper treatment of five categories of variants: (1) orthographic (head word vs. headword), (2) morphological (man servant vs. men servants), (3) syntactic (birth date vs. birth of date), (4) semantic (hereditary disease vs. genetic disease) and (5) pragmatic (Prime minister vs. he). The variability of MWEs is another challenge to knowledge-poor methods, since basic techniques such as lemmatization or stemming of all corpus words, result in overgeneralizations (e.g. customs office vs. custom office) or in overlooking of exceptions (e.g. passersby).

2.4 Why handle MWEs in NLP systems?

It becomes critical to positively master the treatment of MWEs on NLP applications to reduce the inherent impact of producing an ungrammatical or unnatural output. Some NLP systems and tasks in accordance with [Ramisch, 2015] are shown below:

- **Computer-aided lexicography.** Constructing a MWEs dictionary is highly complex and requires more effort than building a lexical resource (e.g. dictionary). It is important that lexical resources start the inclusion of the MWEs in their content. By doing so, humans and machines will be able to process MWE.
- **Optical Character Recognition (OCR).** MWEs could help improve OCR technology, overcoming length limitations of n-gram language models
- **Information Retrieval Systems (IR).** If MWEs are indexed as lexical and semantic units, the accuracy and usefulness –on retrieving the relevant documents to the user– of the system will probably improve dramatically.
- **Foreign Language Learning.** MWEs play an important role in the design of computer environments for foreign language e-learning platforms. In addition, language certificate platforms such as ETS TOEFL can use automatic MWE processing to spot errors and evaluate fluidity of non-native text.
- **Computational Semantics and Machine Translation (MT).** Adding MWEs to MT strategies can greatly improve translation quality, MWEs

are as well considered a challenge in automatic word alignment of two parallel texts.

3 Dealing with Spanish MWEs

The current necessity of correctly identifying and working with multi-word expressions extends to all languages, and particularly in Spanish the challenge can increase if we consider the different dialects spoken among countries and even among regions of the same country. Following the examples presented on the first section of this document, one phrase can be a MWE in Costa Rica, but not be one in Mexico, while Perú and Colombia can use the same MWEs to express different sentiments. Also, there are expressions such that if we combine the possible inflections and the nearly unlimited additions that are allowed the variations of the same MWE are over a thousand. One example of this is '*hacerse el loco*' (#5) that literally means to become crazy, but it's used to express the situation when you witness something and pretend not to see it. The verb '*hacerse*' can inflect in many ways, and '*el loco*' (the crazy) can inflect in gender and number, but both components have to agree on both features and they also need to agree in number with the verb.

There are other situations where changing or removing one word of the MWE can change the meaning. For example, in the expression shown in section 1.1 '*hablar pura paja*'(#20) (small talk), if we remove the verb '*hablar*' leaving only '*pura paja*' it becomes another MWE in Costa Rica that is used to describe persons that say they will do something but never do. Furthermore, there are MWE where the number of possible meanings is very high, even within a single dialect. For example, the phrase '*pura vida*' (#59) in Costa Rica is used to say hello, say goodbye, describe a person who is very nice and easy going, express how you are doing (e.g. '*–hey! how are you? –Pura vida!*'), ask someone how he/she is doing ('*–pura vida? –pura vida!*'), express that a situation is good and exiting, and so on.

These sort of scenarios are constantly present when two Spanish speakers interact. While in general all the meanings can be understood or at least inferred from the context, there can be occasions, specially if they come from different countries, when one speaker might need to ask what the other is referring to or if his intensions are the same as they were understood. If we move to written language, then such clarifications are not possible and inferences could be incorrect.

Moreover, currently most of the existing NLP applications have been developed for English, and although some of them have provided good results when dealing with MWEs, the complexity of these expressions in Spanish grows with the richness of the morphology of the language itself. Therefore, current tools don't have the desired outcome when used directly on Spanish expressions and more resources are needed.

To help with this challenge, we propose a schema for the storage of MWEs for different Spanish dialects, and together with it we provide a corpus of documents where these expressions can be found. The richness of this corpus lies in the fact that it's possible to find the MWE in the meaning as it was intended or in a compositional or literal meaning. This way it's possible to compare the different uses of these expressions and in the future it can be either used, for example, for recognising the original dialect of a certain text based on the MWEs contained in it.

4 Schema for representing Spanish MWEs

The variability and number of MWEs in a single Spanish dialect is very high, now if we consider different dialects at the same time, then the combinations that can be done between expressions, meanings and allowed inflections are massive. To be able to create a schema for the representation of all these variations we first defined a set of morphological, syntactic and semantic properties that allow us to express the different behaviors of the MWEs. Then we defined a high-level model that included all these properties and how the different elements within a MWE relate to them, later based on that we created an XML schema for storing the MWEs. In this section we first describe the properties chosen, then we move to the model description and finally the schema definition.

4.1 Linguistic properties of Spanish MWE in different dialects

The properties chosen to be included in our model are the following:

- **Partial inflection:** describes whether the MWE allows a subset of inflections in its components while preserving its meaning for the dialect.
- **Inflection degree:** used to express three levels of inflection: frozen, semi-inflection or total.
- **Lexical fixedness:** states that if we replace any of the constituents of a MWE by a semantically (and syntactically) similar word, this results in an invalid or literal meaning.
- **Frozen form:** specifies if the constituents of the MWE can only appear in one fixed form.
- **Variety:** encompasses the different part-of-speech categories that a MWE can contain, for example: verb, noun, adjective, adverb, conjunction, preposition, pronoun. For each of them a set of possible inflections is defined

- **Modification:** it indicates whether the MWE allows additions or substitution of words.
- **Compositionality:** it's used to express if a relation between the meaning of the components and the meaning of the MWE exists.
- **Language register:** refers to the level of formality of the MWE. For example, colloquial and casual.
- **Passivization:** indicates whether the MWE can be expressed in passive form while preserving its meaning.
- **Dialect:** indicates to which Spanish dialect this MWE belongs.
- **Meaning:** describes what a person really wants to say when using the multiword expression.

4.2 Model for representing MWEs

Following our goal of creating a schema for Spanish MWEs in different dialects that let us represent the general expression's meaning(s) and properties while also allowing the input of specific variations available in one or more dialects, but not in all of them; we first defined the model on Figure 1 at a high abstraction level. This allowed us to identify which of the properties listed on the previous section applied to the entire expression, which ones could be applied at a word level, which ones were common for all dialects and which ones could take values that vary between the dialects.

For example, the *partial inflection* property (denoted as *allowsInflections* attribute in Figure 1) applies at a MWE level, but at a word level it's also possible to define if a particular token allows inflections or not. *Lexical fixedness* (*allowsSubstitutions* attribute) is similar in the sense that it applies at a MWE level, but on each word it's possible to specify if it can be replaced. On the other hand, *compositionality* works only at a MWE level as well as *passivization* and *language register*. On the other hand, *variety* property referring to the morphological analysis is specified only at each token.

As it's shown in Figure 1 the main component is the MWE, which relates to other entities that have the purpose of describing that MWE. For example, we defined a set of properties for the overall MWE and then for the *Token* entity a new set of properties is listed. This targets the mentioned situation where the overall MWE allows, for example, inflections, but not on every token. So in this case, the MWE property will have a value of *true*, but each token will also detail if it allows inflections or not. If nothing is specified at a token level, then it is assumed that it inherits the property value from the parent MWE component.

The path entity is envisioned as a sequence of tokens that can be part of the base, additional or substitute token lists. The idea is to be able to depict the

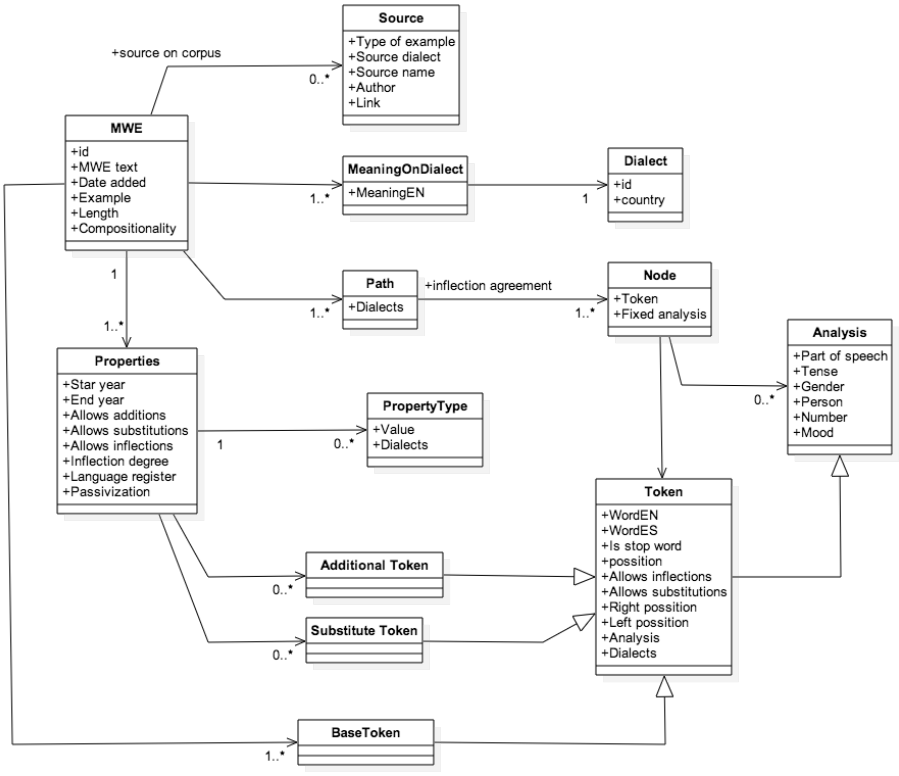


Figure 1: High level model for representing MWE

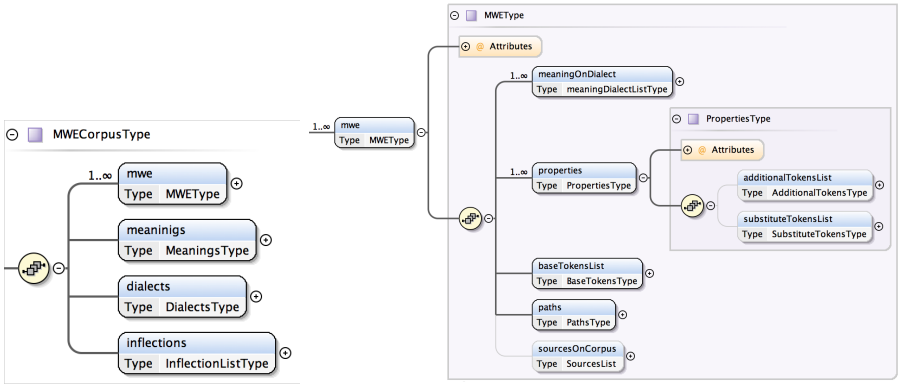
different ways a single MWE can be expressed, as well as any restriction –in terms of agreement on inflection categories– that can exist between the tokens of a MWE. An example of this will be explained with more detail in the next section.

4.3 Schema description

The schema presented in this section is based on the previously described model and it represents a MWE as an item that encompasses the morphological, syntactic and semantic properties defined in section 4.1. In order to simplify the content of the MWE expression we defined the dialects, meanings and inflections outside of the MWEs, so that whenever an expression is linked to any of those elements it only needs to reference the corresponding identifier. The higher level structure on the schema is shown on Figure 2a.

Then, as can be seen in Figure 2b, for each MWE we defined:

- *meaningInDialect*, which is a list of meanings for the expression and the corresponding dialect where that meaning is used. The purpose of this element is to tackle the situations when a single MWE can have different meanings. Also, since it references the dialects and meanings defined at



(a) Main structure (b) MWE expression element

Figure 2: High level structures of the XML schema

the higher level, it can be used to identify MWEs that are different in terms of their constituent elements, but have the same meaning. This can be done by querying the meaning identifier and get the list of associated MWEs.

- a *properties* element that includes the semantic, syntactic and morphological properties that apply to the overall expression. It also holds optional lists for the allowed additional or substitute tokens.
- *baseTokensList* that has the list of words (referred as tokens from here on) that constitute the base form of the MWE.
- a *paths* element that is used to detail if there are any tokens in the MWE that must agree in one or more inflection categories. It also helps to specify if it's possible to skip one of the base tokens without losing the MWE meaning, resulting in an alternate order of the base tokens, hence creating another path for the expression.
- *sourcesInCorpus* holds a list of references to the documents in the corpus that contain this MWE. It also includes an attribute to describe if the meaning of the MWE in that particular text is positive (expected idiomatic meaning), neutral (compositional meaning) or negative (the components of the MWE occur in the text but not in one appropriate syntactic group).

The three lists of tokens (base, additional and substitute) that can be associated with a MWE are all composed of an element of type *token*. The definition of this type is in Figure 3. In our proposal it is conceived as the most basic element of a MWE, it holds the word on a particular position and it's described with a set of attributes that define: the morphological analysis, if the particular word represented is a stop word, if it allows inflections or substitutions, the associated dialect and the left and right positions of the adjacent tokens.

These last two attributes are used when the token is part of a list of additional of substitute words to be able to specify exactly where the element should be placed. For example, in the expression '*aventar la madre*' (#58) which means '*to insult someone*' it's possible to substitute the verb '*aventar*' ('*throw*) in the first position with the verb '*mentar*' ('*mention*'), resulting in a new form of the MWE: '*mentar la madre*'.

On the other hand, the dialect property is set in case the token is only allowed in some but not all of the dialects. It's important to highlight that the schema was designed in a way that, unless it's strictly specified, all properties and elements within a *mwe* tag are valid in all of the languages defined in the *meaningInDialect* list. That is, if one MWE has a list of, for example, substitute or additional tokens without any *dialect* property, then it can be understood as if that particular token is accepted in all dialects mentioned in *meaningInDialect*. This is the case of the example in Figure 4 where the substitute token '*mentar*' doesn't have the *dialect* property and is therefore assumed valid in both Costa Rica and Perú.

```
<xs:complexType name="TokenType">
    <xs:annotation> [7 lines]
    <xs:sequence>
        <xs:element name="wordES" type="xs:token"/> <!-- word in spanish -->
        <xs:element name="wordEN" type="xs:token" /> <!-- literal translation to english -->
    </xs:sequence>
    <xs:attribute name="id" type="xs:ID" use="required" > [4 lines]
    <xs:attribute name="isStopWord" type="xs:boolean" use="required"> [6 lines]
    <xs:attribute name="possition" type="xs:integer" use="required"  />
    <xs:attribute name="allowsInflections" type="xs:boolean" use="required" />
    <xs:attribute name="allowsSubstitutions" type="xs:boolean" use="required" />
    <xs:attribute name="rightPossition" type="xs:integer" use="optional" />
    <xs:attribute name="leftPossition" type="xs:integer" use="optional" />
    <xs:attribute name="analysis" type="xs:IDREF" use="required" />
    <xs:attribute name="dialects" type="xs:IDREFS" use="optional" />
    <xs:attribute name="originalLanguage" type="xs:string" use="optional" />
</xs:complexType>
```

Figure 3: Definition of a Token within a MWE

An example of how these properties and configurations are expressed in the schema can be found in Figure 4. There we can see that for this MWE the allowed substitutions is a closed list composed of only the verb '*mentar*', therefore, it's explicitly mentioned in the schema in lines 10 to 16 of the image. As for the option of adding new tokens, the property *allowsAdditions="false"* (line 5) states that it is not possible for this MWE, hence the list of additional tokens is omitted. Also, the second and third tokens don't allow either substitutions (*allowsSubstitutions="false"* on lines 25 and 30) nor inflections (*allowsInflections="false"* on lines 24 and 29). These properties on the token are used to describe at a more detailed granularity the changes that a MWE can undergo, given that the same properties at the MWE level are all set to "true" (line 4) because they apply to the complete expression.

Following the remaining lines on Figure 4 we can see how *paths* element is used to describe (1) how the tokens of a MWE can be ordered taking into account substitutions and additions, (2) if some of the tokens have a fixed form and what it is by means of the *fixedAnalysis* attribute, and (3) if there

```
<mwe id="MWE10" mweText="aventar la madre" length="3">
    <meaningOnDialect id="MWE10ISCR" meaning="IS" dialect="CR"/>
    <meaningOnDialect id="MWE10ISPE" meaning="IS" dialect="PE"/>
    <properties>
        <allowsAdditions value="false" />
        <allowsSubstitutions value="true"/>
        <allowsInflections value="true"/>
        <languageRegister value="Vulgar"/>
        <passivization value="true" />
        <substituteTokensList>
            <substituteToken id="MWE10_ATkn1" isStopWord="false" possition="1"
                allowsInflections="true" allowsSubstitutions="true" analysis="V.W">
                <wordES>mentar</wordES>
                <wordEN>mention</wordEN>
            </substituteToken>
        </substituteTokensList>
    </properties>
    <baseTokensList id="MWE10_BTkns">
        <baseToken id="MWE10_BTkn1" isStopWord="false" possition="1" allowsInflections="true"
            allowsSubstitutions="true" analysis="V.W">
            <wordES>aventar</wordES>
            <wordEN>winnow</wordEN>
        </baseToken>
        <baseToken id="MWE10_BTkn2" isStopWord="true" possition="2" allowsInflections="false"
            allowsSubstitutions="false" analysis="DET.fs">
            <wordES>la</wordES>
            <wordEN>the</wordEN>
        </baseToken>
        <baseToken id="MWE10_BTkn3" isStopWord="false" possition="3" allowsInflections="false"
            allowsSubstitutions="false" analysis="N.fs">
            <wordES>madre</wordES>
            <wordEN>mother</wordEN>
        </baseToken>
    </baseTokensList>
    <paths>
        <path>
            <node token="MWE10_BTkn1"/>
            <node token="MWE10_BTkn2" fixedAnalysis="DET.fs">
                <inflectionAgreement gender="$g"/>
                <inflectionAgreement number="$n"/>
            </node>
            <node token="MWE10_BTkn3" fixedAnalysis="N.fs">
                <inflectionAgreement gender="$g"/>
                <inflectionAgreement number="$n"/>
            </node>
        </path>
        <path>
            <node token="MWE10_ATkn1"/>
            <node token="MWE10_BTkn2" fixedAnalysis="DET.fs">
                <inflectionAgreement gender="$g"/>
                <inflectionAgreement number="$n"/>
            </node>
            <node token="MWE10_BTkn3" fixedAnalysis="N.fs">
                <inflectionAgreement gender="$g"/>
                <inflectionAgreement number="$n"/>
            </node>
        </path>
    </paths>
</mwe>
```

Figure 4: First example of a MWE in the proposed schema

are tokens that must agree in one or more inflection categories For example, lines 38, 42, 49 and 53 on Figure 4 specify that tokens 2 and 3 must always be, respectively, a feminine singular determiner (DET.fs) and a feminine singular noun (N.fs). For the first tokens, either the base one or the substitute no inflection is specified because all of them are allowed.

The third quality mentioned above that can be expressed on a path is the agreement requirements between some tokens of the MWE. This is seen more clearly on Figure 5. In this case the MWE is '*hacerse el loco*' and like was described in section 3, all the base tokens allow inflections. However, the *number* between the three tokens must agree and the *gender* of the second and third tokens must also be the same. This restriction is expressed in the schema by the introduction of unification variables that aim to express that whatever inflected form is used on token 3 for gender ($g) they must be the same on

Token 4, and that the number ($n) on all three tokens must also be the same. Note that since the *fixedAnalysis* attribute was set as optional it was omitted from this example, implying that all inflections for this token are allowed.

Figure 5: Second example of a MWE in the proposed schema

Another property defined in the schema that is important to highlight is the *notFollowedBySpace* attribute for tokens under a path. The aim of this boolean value is to detail if the current token includes a space before the next token of the path. The reason for this attribute is the type of MWE that have the reflexive pronoun '*se*'. For example, in Figure 5 the MWE '*hacerse el loco*' doesn't have a space between the first token ('*hacer*') and the second ('*se*'), but if we change the order and add a space, while still keeping the meaning, the phrase becomes '*se hace el loco*'. These two possible ways of using the MWE are shown inside the paths tag of the XML of Figure 5 (lines 34-73).

The glossing rules of the *analysis* elements are based on [Bickel et al., 2008]. Hence, when an *analysis* is specified within a MWE, it references an instance

already defined in the higher level of the schema that describes the properties of that component, including part-of-speech class, gender, type, number, person, tense and mood. Figure 6a shows an extract of the morphological analysis definition on the schema file, while 6b has an extract of the list of analysis as defined in the XML document.

```
<xs:complexType name="AnalysisType">
    <xs:annotation> [8 lines]
    <xs:attribute name="id" type="xs:ID" use="optional" /> <!-- possible example VJ3s -->
    <xs:attribute name="partOfSpeech" type="POSType" use="optional" > [8 lines]
    <xs:attribute name="tense" type="TenseType" use="optional" > [8 lines]
    <xs:attribute name="gender" type="GenderType" use="optional" > [6 lines]
    <xs:attribute name="person" type="PersonType" use="optional" > [6 lines]
    <xs:attribute name="number" type="NumberType" use="optional" > [6 lines]
    <xs:attribute name="mood" type="MoodType" use="optional" > [6 lines]
</xs:complexType>
<xs:simpleType name="POSType"> <!-- Part-Of-Speech -->
    <xs:restriction base="xs:token">
        <xs:enumeration value="V"/> <!-- verb -->
        <xs:enumeration value="DET"/> <!-- determiner -->
        <xs:enumeration value="N"/> <!-- noun -->
        <xs:enumeration value="PREP"/> <!-- preposition -->
        <xs:enumeration value="A"/> <!-- adjective -->
        <xs:enumeration value="ADV"/> <!-- adverb -->
        <xs:enumeration value="CNJ"/> <!-- conjunction -->
        <xs:enumeration value="PRO"/> <!-- pronoun -->
        <xs:enumeration value="PREPDET"/> <!-- preposition + determiner, e.g. del, al -->
    </xs:restriction>
</xs:simpleType>
<xs:simpleType name="GenderType">
    <xs:restriction base="xs:token">
        <xs:enumeration value="m"/> <!-- masculine -->
        <xs:enumeration value="f"/> <!-- feminine -->
        <xs:enumeration value="n"/> <!-- neuter -->
        <xs:enumeration value="$g"/> <!-- unification value -->
    </xs:restriction>
</xs:simpleType>
<xs:simpleType name="NumberType"> [6 lines]
<xs:simpleType name="PersonType"> [7 lines]
<xs:simpleType name="TenseType"> [15 lines]
<xs:simpleType name="MoodType"> [6 lines]
```

(a) Extract of analysis definitions

```
<inflections>
    <inflection id="N.fs" partOfSpeech="N" gender="f" number="s"/>
    <inflection id="N.fp" partOfSpeech="N" gender="f" number="p"/>
    <inflection id="N.ms" partOfSpeech="N" gender="m" number="s"/>
    <inflection id="N.mp" partOfSpeech="N" gender="m" number="p"/>
    <inflection id="PRO.3s" partOfSpeech="PRO" person="3" number="s"/>
    <inflection id="DET.fs" partOfSpeech="DET" gender="f" number="s"/>
    <inflection id="DET.fp" partOfSpeech="DET" gender="f" number="p"/>
    <inflection id="DET.ms" partOfSpeech="DET" gender="m" number="s"/>
    <inflection id="DET.mp" partOfSpeech="DET" gender="m" number="p"/>
    <inflection id="ADV" partOfSpeech="ADV"/>
    <inflection id="A.fs" partOfSpeech="A" gender="f" number="s"/>
    <inflection id="A.fp" partOfSpeech="A" gender="f" number="p"/>
    <inflection id="A.ms" partOfSpeech="A" gender="m" number="s"/>
    <inflection id="A.mp" partOfSpeech="A" gender="m" number="p"/>
    <inflection id="PREPDET.ms" partOfSpeech="PREPDET" gender="m" number="s"/>
    <inflection id="V.W" partOfSpeech="V" tense="W"/>
    <inflection id="V.J1s" partOfSpeech="V" tense="J" person="1" number="s"/>
    <inflection id="V.J1p" partOfSpeech="V" tense="J" person="1" number="p"/>
    <inflection id="V.P1s" partOfSpeech="V" tense="P" person="1" number="s"/>
    <inflection id="V.P1p" partOfSpeech="V" tense="P" person="1" number="p"/>
    <inflection id="V.J2s" partOfSpeech="V" tense="J" person="2" number="s"/>
    <inflection id="V.J2p" partOfSpeech="V" tense="J" person="2" number="p"/>
```

(b) Extract of analysis instances

Figure 6: Examples of morphological analysis definition and instances in the XSD and XML documents

In a similar way we defined the meanings for the MWEs, such that when an element *meaningInDialect* is instantiated it refers to the meaning definition stated at a global level. For example, lines 2-5 of Figure 5 point to the meaning and dialects shown in Figure 7

```
<dialects>
    <dialect id="CO" country="Colombia"/>
    <dialect id="CR" country="Costa Rica"/>
    <dialect id="MEX" country="Mexico"/>
    <dialect id="PE" country="Peru"/>
</dialects>
<meaninigs>
    <meaning id="TD" meaning="To Die"/>
    <meaning id="BP" meaning="A big problem started"/>
    <meaning id="SU" meaning="To screw up something"/>
    <meaning id="WP" meaning="To witness something and pretend not to see it"/>
    <meaning id="TL" meaning="To take the lead"/>
    <meaning id="DF" meaning="To do things faster"/>
    <meaning id="DB" meaning="To do things better"/>
    <meaning id="GT" meaning="To get one's act together"/>
    <meaning id="BG" meaning="Bad people have good luck"/>
    <meaning id="GU" meaning="To give up"/>
    <meaning id="IS" meaning="To insult"/>
    <meaning id="CP" meaning="To cheat your partner"/>
</meaninigs>
```

Figure 7: Examples of dialects and meanings

5 A method to construct a web-based corpus to extract MWE candidate examples aimed by a crowd-sourcing human interpretation

As a start point of this section, let's recall some important concepts related to corpus in literature. In [Ramisch, 2015] a corpus is defined as body of texts used in empirical language studies, usually employed to represent the target language, where the meaning of represent depends on the context (e.g. application, domain, genre, sub-language). Usually a corpus is structured as a set of documents, each document being composed of several paragraphs, which in turn are sequences of sentences.

A corpus that contains data in one language is called monolingual corpus, on the other hand, a corpus in several languages is called multilingual corpus, when the sentences in one language are translations of sentences to another language, we consider it as a sentence-aligned parallel corpus. Finally depending on the information that the corpus holds, a corpus can be categorised as general-purpose to refer to corpora that contain a wide variety of texts corresponding to most common language use over a given time span, or specialised to allude a corpus containing texts of a specific knowledge domain or sub-language, like botany, computer science or sailing.

5.1 Corpus characteristics

There are some important characteristics of a corpus that could impact the evaluation results for the intended reason that it was constructed for. Those significant characteristics are: (1) size – the larger the corpora, the bigger chances to obtain more MWE candidates–, (2) nature – domain and genre of the texts – and (3) level of linguistic analysis used.

5.2 Web as corpus

According to [Boos et al., 2014] during the past years, initiatives for constructing very large corpora have been increasing especially using the Web as corpus. This approach employs crawlers to collect sets of texts which are subsequently cleaned – to extract only the textual contents – and filtered – to remove duplicate material and noise.

Different methods to collect corpora from the web are available, for example corpus extraction by using focused crawling, starting from a domain specific page guiding the crawler based on the proportion of relevant words that can be found in the text of that page and of all the other pages that belong to the same host. Likewise parallel corpora could be created by empowering the crawlers to identify equivalent pages in multiple languages, i.e using Wikipedia as a source for parallel corpora to construct aligned sentences. Without doubt these web

building corpus strategies represent an inexpensive way of constructing a large resource, especially for languages for which freely available resources of this magnitude are limited and insufficient.

For this work we designed a method based on some steps of the WaCKy [3](Web-As-Corpus Kool Yinitiative) methodology, adding an automatic process based on SQL code, database quality tools and techniques (Oracle Enterprise Data Quality [4], to basically normalise the data and tokenize each MWEs), for constructing our Spanish corpus. In the following section we explain in detail the entire process we build to construct our corpus.

5.3 Constructing the corpus

The method we designed consists in basically 5 steps shown in Figure 8. Each stage is explained in detail in the following subsections. In order to give a complete outlook of the entire process we include in Figure 8 a 6th step that is going to be explained separately in the subsection *Crowd-Sourcing*.

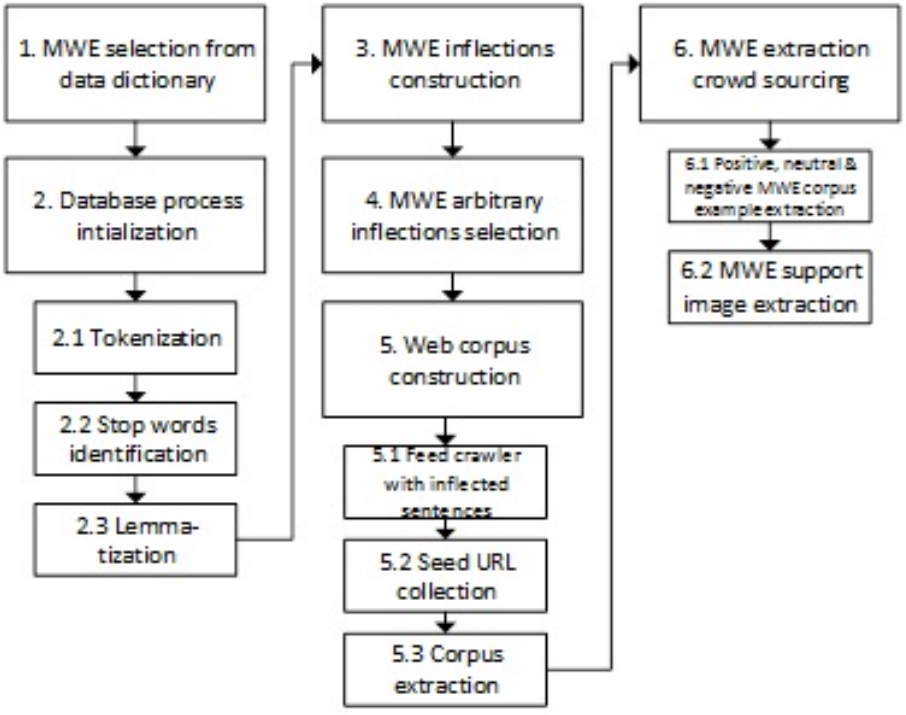


Figure 8: Designed method to construct a web based corpus

[3] `http://wacky.sslmit.unibo.it/doku.php?id=start`
[4] http://www.oracle.com/technetwork/middleware/oedq/overview/index.html

5.3.1 MWE selection from data dictionary

The first step of the above mentioned process is the selection of a set of MWEs from our data dictionary. Each time that we ran the entire process from steps 1 to 5, we selected a set of 50 MWEs. The selection process was purely random assuring the inclusion of MWE examples from the different dialects that our current study covers.

5.3.2 Database process initialization.

In this stage of the process the data preparation tasks are performed.

I **Tokenization**
We started by applying a tokenization activity on each MWE, this process was performed in combination with Oracle EDQ tool and using a PL/SQL code with a regular expression (REGEXP$_S$UBSTR) function that decomposes the MWE in tokens (words). An example of this process is shown in the figure 9.

Figure 9: Tokenization process example

II **Stop word identification**
In this step we used a SQL process to identify stop words (words that present no significant relevance for searching). We took into consideration two online Spanish[5] databases, containing a vast list of stop words, and we applied the stop word identification function to all the tokens for the given MWE. Figure 10 illustrates graphically this process.

[5] `http://www.ranks.nl/stopwords/spanish`, `https://code.google.com/p/stop-words/`

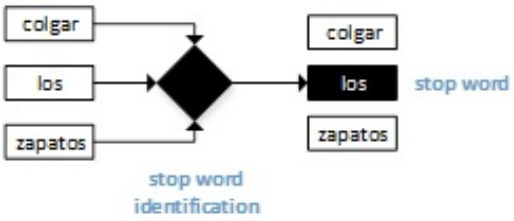


Figure 10: Stop word identification process

III **Lemmatization**
Just after the stop word identification we used the AGME database [6] and a PL/SQL function to identify the lemma of the given word (token) to retrieve all possible inflections of it. Figure 11 shows a graphical example of this process. It is important to stress out that in this activity we exclude the stop word identified in the previous step.

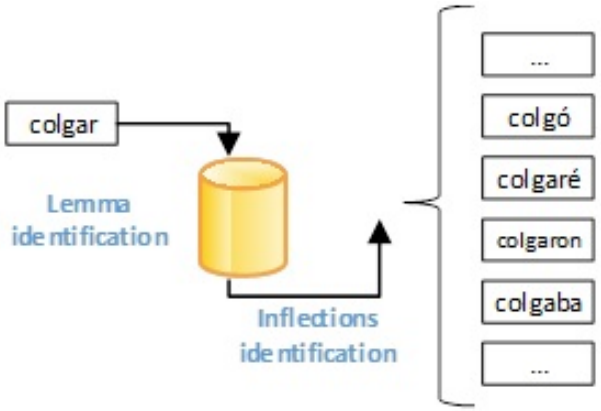


Figure 11: Lemmatization and inflections identification processes

5.3.3 MWE inflections construction

Once all token inflections were identified, we ran a PL/SQL code we developed in order to formulate new sequences having the same structure of the MWE, but with all the possible inflections obtained in the previous stage. Just as a remark stop words that were identified, remain in the same grammar state as they are in the original MWE. The output of this process was huge, the Table 2 shows the 3 MWEs with the biggest quantity of inflected sequences created.

[6] `http://www.cic.ipn.mx/~sidorov/agme/`. Automatic Morphological analysis of Spanish database. It's based on the FreeLing Database, contains around 26,000 Spanish lemmas, including clitic words

Table 2: Top 3 of MWE with the biggest quantity of inflections created

Nr	MWE	Inflected sentences
1	Quedarse con el ojo cuajado	142578
2	Comer callado	71823
3	Ponerse las pilas	71022

5.3.4 MWE inflections arbitrary selection

To finish with the data preparation part, due to the limitations of the web crawler that we used for our study – mentioned later in this paper – we needed to find a way to create an extract of the inflected sequences in order to be the input of the crawler. To accomplish this task we developed a PL/SQL procedure that performs a random choice of the existing sentences and creates and extract of around 100 sentences per MWE, by using the package `dbms_random.value`.

5.3.5 Web corpus construction

Once we concluded the data preparation part of our method, we continued with the web corpus construction. To accomplish this task we mainly used the BootCat[7] web crawler. This crawler is an open source tool specialised for text extraction from web. In the following subsections we explain the process followed.

I **Feed Crawler with inflections obtained**
As we said in the last step of the data preparation, we randomised a selection of inflections based on the given MWE; we used this extract of inflections as the input for the crawler. To perform the search of possible URL containing text with the target words.

II **Seed URL collection**
In addition to inflections, an optional attribute that could be defined to the crawler is the specification of the exact URL/domains from which we want to get the text data /corpus. We defined some specific URLs based on the dialect of the given MWE in order to force somehow the crawler to search first in those sites. Nonetheless, we didn't restrict the process to search in other websites because we were looking to obtain as much information we could.
It is important to mention that BootCat uses the BING Search API to perform the URLS search on the web. The free (without cost) usage of this API is restricted to 5,000 searches/transactions per month, and it is

[7] `http://bootcat.sslmit.unibo.it/`

necessary to create an account in its corresponding website to start using it.

III **Corpus extraction**
The last step of the corpus construction section, is fundamentally the corpus extraction process, this activity occurs automatically and is performed by the web crawler using the specified parameters. As an output of the process the web crawler creates a txt file containing the appended text from all the different websites where the data was crawled, it includes specific tags that allows the user to identify where the data from a specific website starts.
The webcrawler in combination with the BING search API[8] use a NEAR function to look for the data. For example, if we define to look up for text having the compound collocation "Computer Science", it is performing the search by employing NEAR function which means that will look for the term "Computer Science" but it will not restrict to search for the fixed term, it will also fetch the text having in a near position from one term to another, for example: "Computer and Information Technologies Sciences;' and 'Computer and telecommunications how they help in Biomedical Science'.

6 MWE example extraction from the created corpus

As we explained in the subsection 5.3, we added a 6th step to our entire corpus creation process, that although is not precisely an activity involved in corpus creation, we consider it the most important one due the human interaction need. This step falls on the task of analysing the created corpus to extract a MWE example from it. For this purpose we developed a user friendly web form with the main purpose of serving as a point for collecting MWE examples from the corpus. This approach is aimed by the crowdsourcing strategy, a really good collaborative methodology in which we invited users to participate as volunteers in the MWE examples from corpus extraction process. The Figure 12 illustrates graphically this crowdsourcing strategy used.

In the appendix section of this document we list the resources created used in this crowd sourcing methodology. (web form, MWEs dictionary, corpus repository and tutorials)

6.1 Positive, Neutral and Negative MWEs examples

For the extraction of the MWEs examples, we defined three types of possible MWEs examples: positive, when the MWE example found has the same id-

[8] `https://datamarket.azure.com/dataset/bing/search`

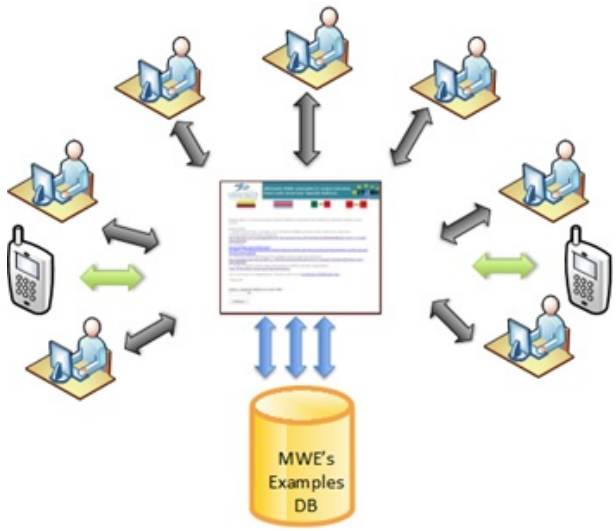


Figure 12: MWE extraction based on crowdsourcing strategy

iomatic meaning of the MWE, neutral: when the MWE example found has a literal meaning in the text, and negative: when the components of the MWE occur in the text but not in one appropriate syntactic group. Figure 13 illustrates a negative example found regarding the MWE "Colgar los zapatos"

6.2 MWEs support image extraction

It is well known that a picture is worth a thousand words, hence to support the MWEs understanding that our study covers, we attempted to create the basis of an image database to relate an illustration to the MWE example. Figure 7 presents 3 pictures related with the 3 types of examples gathered for the Mexican MWEs "Colgar los zapatos".

Figure 13: Image support for MWE "Colgar los zapatos": A) represents the positive meaning that is "to die", B) refers to the neutral example that is the literal meaning "to hang shoes" and C) represents the negative example found that includes the concepts of "hang" and "shoes" but are not conforming a MWE.

The creation of an image database is a time consuming activity, therefore this task will be treated as a different project to continue in future work related

Table 3: The distribution of MWE per Spanish dialect in the database

Spanish dialect	MWEs database
Colombian	14
Costa Rican	19
Mexican	10
Peruvian	27

to our study.

7 MWE database

Following the XML schema defined above and including the positive, negative and neutral examples found in the corpus extraction, we created a MWE database. We started by listing 256 MWEs from 4 different Spanish dialects: 108 from Mexico, 57 from Colombia, 43 from Costa Rica and 48 from Peru. Out of those we generate the corpus for 110 and finally we created the XML database for 40 Spanish MWE linked to 43 different meanings, all having a link to the respective example(s) in the corpus.

The access to all this resources can be found in the Appendix. In the next section we will discuss the limitations and the evaluation of our work.

8 Limitations and evaluation of resources

This part of our document is intended to express some interdictions identified during the elaboration of the MWEs database and MWEs corpus for examples extraction, to take them in consideration during the evaluation of the resources and to propose a future method that diminishes the impact of these limitations on the produced resources.

8.1 MWEs database shortcomings

As we have explained so far, the vast variety of inflections that an idiomatic MWE could have in Spanish dialects is enormous. To have an idea, in Table 2 we pointed out some quantities of automatic inflected sequences created for one given MWE. To make the situation more complex, this number will increase if the analysis of the context where a given MWE appears is taking in consideration, for example, lets analyse the MWE “Hacerse el loco” whose meaning is to pretend to be crazy, or to pretend not to know something, apart from the possible grammar inflections that the 3 words could have “Hacerse”,“el”

and “loco” the context plays an important role in this example, Why? Please have a look on the following Spanish sentence:

“Ella se hizo la loca, cuando preguntaron quién había tomado el dinero” (She pretend not to know, when they asked who had taken the money)

Now, please refer to the Figure 14, that shows the decomposition of the MWE “Ella se hizo la loca”

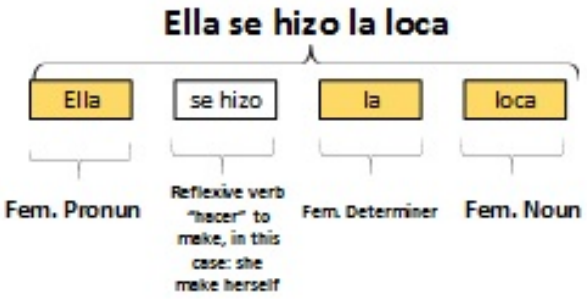


Figure 14: MWE Glossing Rules

As we see in the figure above, each yellow square must be in concordance of gender and number to make the sentence semantically correct. This analysis of context to produce semantically correct inflected sequence of a given MWE is a challenging activity that due to the huge of interminable possible contexts in which that MWE could appear, was not taken in consideration when building the DB.

8.2 MWEs corpus creation for examples extraction

The web based corpus we built whose construction method is explained in section 5 of this study, used different technology tools to perform the search and extraction of web text. Although these tools are really good approaches for an inexpensive corpus generation based on the web, we identified some restrictions on them that could impact the results we produced. The coming paragraphs explain in summary those limitations.

In order to build a well oriented corpus, we defined a set of specific URLs from the different countries that our study covers, those websites (rich in meaningful and important information) were mainly, online newspapers, new trends portals, important blogs – to mention few –, in which the possibilities to find an example of a specific MWE area high.

After building this URLs list, we fed the web crawler with it and we executed the process. After seeing the results we noticed that majority of the websites we defined were discarded, in other words, no data was extracted from them during the automatic process. Digging in this situation, we found that most of the websites defined use different rules in their code to not allow the web

crawlers to extract data, and although when we performed manual searches against those websites we found good MWE example, we were not able to bring them to our automatic corpus creation due to this constraint found.

Additionally, as explained in section 5, to perform the searches the web crawler uses the BING Search API available for free usage, allowing to perform 5000 searches (transactions per month). Therefore, we had to restrict the searches of the inflected sequences of a given MWE, by choosing a random set of 100 sentences per each MWE. This limitation reduced the margin of finding MWE example candidates in the created corpus.

9 Conclusions and Future Work

In this research we have compiled a list of properties that help describe Spanish multi-word expressions from a morphological, syntactic and semantic point of view. Furthermore, we have identified a set of relationships between the components of the MWE themselves and with the particular set of dialects (Colombia, Costa Rica, Mexico and Perú), to help us profile all the variations a single expression can undergo when it's used in different Spanish dialects. Based on this analysis we developed a schema for the general representation of MWEs, while at the same time allowing the specification of the particularities they have in every country they are used. Moreover, we described a process for the creation of a corpus for these expressions using web crawling and crowd sourcing techniques.

This work can be extended in different ways, first of all by increasing both the MWE database and the corpus, then it would be possible to measure the effectiveness and quality of the corpus in terms of the number of MWE expressions that it contains and how many examples of each of them can be found. For the extension of the corpus it would also be possible to link the web crawling phase to the MWE database in order to consider the fixedness of some MWEs and reduce the search space to only those websites that meet the restrictions.

Another path that can be followed is to build a parallel corpus for the MWEs that have an equivalent in other languages, allowing non-native speakers to understand in context what is meant with the expression. Also, a search engine for MWE can be built on top of the corpus and the database to allow people to look for a MWE and obtain websites where it's used with its idiomatic meaning. This can be useful because currently a large number of web pages that have MWEs are actually explaining the meaning of those phrases and not using them in context. However, with our filtered corpus it'd be possible to obtain actual texts that use MWEs and –if needed– get the meaning and other properties from the DB.

Furthermore, we need to consider the fact that the number of MWEs used nowadays continues to grow, specially thanks to social networks and video-sharing websites like *YouTube*, where a video with a catchy phrase can be so popular that an entire country relates to it and becomes part of the colloquial register. Hence, the applications that can be developed with NLP tools that deal with MWE are vast and must be kept up to date. For example, with a bigger corpus and database of MWEs it would be possible to create an application for identifying the country of origin of a particular document considering the MWEs it contains. Another application can be to build a model for machine translation of MWEs from one Spanish dialect to the other. This last work can be very useful considering that per dialect there are many MWEs consisting of just a single word that is usually very specific to a country. Thus, an entire text written with those sort of expressions can be unintelligible for

another Spanish speaker. However, these are only a few of the possible uses that can be derived from this initial work.

Acknowledgements

We would like to thank Professor Grigori Sidorov from Instituto Politécnico Nacional of Mexico for his valuable contribution regarding the usage of AGME database and for his suggestions regarding different ways to construct inflected sequences of a given MWE.

References

[Al-Haj, 2009] Al-Haj, H. (2009). *Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition*. PhD thesis, University of Haifa.

[Attia et al., 2010] Attia, M., Tounsi, L., Pecina, P., van Genabith, J., and Toral, A. (2010). Automatic extraction of arabic multiword expressions.

[Attia, 2006] Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, FinTAL'06, pages 87–98, Berlin, Heidelberg. Springer-Verlag.

[Bickel et al., 2008] Bickel, B., Comrie, B., and Haspelmath, M. (2008). The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February*.

[Boos et al., 2014] Boos, R., Prestes, K., and Villavicencio, A. (2014). Identification of multiword expressions in the brwac. In *Proceedings of LREC*.

[de Caseli et al., 2010] de Caseli, H. M., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

[Gralinski et al., 2010] Gralinski, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational lexicography of multi-word units: How efficient can it be? In *Workshop Multiword Expressions: from Theory to Applications*.

[Grégoire, 2009] Grégoire, N. H. W. (2009). *Untangling Multiword Expressions, A study on the representation and variation of Dutch multiword expressions.*, volume 224. LOT.

[Itai and Wintner, 2008] Itai, A. and Wintner, S. (2008). Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

[Oflazer et al., 2004] Oflazer, K., Say, B., et al. (2004). Integrating morphology with multi-word expression processing in turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71. Association for Computational Linguistics.

[Ramisch, 2015] Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.

[Sag et al., 2002] Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume

2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Appendices

A Multiword Expressions in Spanish

List of Multiword Expressions in Spanish from Colombia, Costa Rica, Mexico and Perú. The glossing rules used in this document are based on `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`

(1) Colgar los zapatos
Colgar.V:s el.DET:mp zapatos.N:mp
Hang.V:s the.DET:mp shoe.N:mp
To hang the shoes
'to die' [MEX]

(2) Colgar las tenis
Colgar.V:s la.DET:fp tenis.N:fp
Hang.V:s the.DET:fp sneaker.N:fp
To hang the sneakers
'to die' [CR]

(3) Estirar la pata
Estirar.V:s la.DET:fs pata.N:fs
Stretch.V:s the.DET:fs leg.N:fs
To stretch the leg
'to die' [COL, CR, MEX, PE]

(4) Ponerse las pilas
Poner.V:REFL la.DET:fp pila.N:fp
Put.V:REFL the.DET:fp battery.N:fp
Put the batteries on yourself
'to get one's act together' [COL]
'to do things in a better way' [CR]
'to be more active' [MEX]
'to do things faster' [PE]

(5) Hacerse el loco
Hacer.V:REFL el.DET:ms loco.N:ms

Become.V:REFL the.DET:ms crazy.N:ms
To act crazy

'to witness something and pretend not to see it' [COL, CR, MEX, PE]

(6) Hacerse de la vista gorda
Hacer.V:REFL de.PREP la.DET:fs vista.N:fs gorda.A:fs
Make.V:REFL of.PREP the.DET sight.N:fs fat.A:fs
To have a heavy sight

'to witness something and pretend not to see it" [COL, CR, MEX, PE]

(7) Por si las moscas
Por.PREP si.CNJ la.DET:fp mosca.N:fp
For.PREP if.CNJ the.DET:fp fly.N:fp
In case there are flies around

'just in case' [COL, CR, MEX, PE]

(8) Estar aguja
Estar.V:s aguja.N:fs
Be.V:s needle.N:fs
To be like a needle

'to be out of money' [PE]

(9) Estar limpio
Estar.V:s limpio.A:ms
Be.V:s clean.A:ms
To be clean

'to be out of money' [CR]

(10) Estar bruja
Estar.V:s bruja.N:fs
Be.V:s witch.A:ms
To be like a witch

'to be out of money' [MEX]

(11) Le amarró el perro
Le.PRON:od amarrar.V:J3s el.DET:ms perro.N:ms
It.PRON:od tie.V:J3s the.DET:ms dog.N:ms
To tie a dog to somebody

'to owe someone money' [CR]

(12) Tener una arruga
Tener.V una.DET:fs arruga.N:fs

Have.V a.DET:fs wrinkle.N:fs
To have a wrinkle

'to owe someone money' [PR]

(13) Hay gato encerrado
Haber.V:P3s gato.N:ms encerrado.V:Kms
Be.V:P3s gato.N:ms encerrado.V:Kms
A cat is locked somewhere

'something fishy' [COL, CR, PE]

(14) Huele a gato encerrado
Oler.V:P3s a.PREP gato.N:ms encerrado.V:Kms
Smell.V:P3s to.PREP gato.N:ms encerrado.V:Kms
Smells like a locked cat

'something fishy' [MEX]

(15) Póngale la firma
Put the signature
Poner.V:Y3s la.DET:fs firma.N:fs
Put.V:Y3s the.DET:fs signature.N:fs
To sign something

'be sure that something will happen' [COL, CR]

(16) Te lo firmo
Te.PRON:od lo.DET:ns firmo.V:P1s
You.PRON:od it.DET:ns sign.V:P1s
To sign something

'be sure that something will happen' [PE]

(17) Hacer la barba
Hacer.V la.DET:fs barba.N:fs
Do.V the.DET:fs beard.N:fs
To shave

'suck up to somebody' [MEX]

(18) Pasar la brocha
Pasar.V la.DET:fs brocha.N:fs
Pass.V the.DET:fs paintbrush.N:fs
To pass a paintbrush over something

'suck up to somebody' [CR]

(19) Tener una laguna mental
Tener.V una.DET:fs laguna.N:fs mental.A:fs

Have.V a.DET:fs lagoon.N:fs mental.A:fs
To have an empty space in your mind
'not able to remember something' [COL, CR, MEX]

(20) Hablar paja
Hablar.V paja.N:fs
Talk.V straw.N:fs
To have straw coming out of your mouth
'small talk' [COL, CR]
'to tell a lie' [CR]

(21) Comer pavo
Comer.V pavo.N:ms
Eat.V turkey.N:ms
To eat turkey
'to go to a party and nobody ask you to dance' [COL]

(22) Parar bolas
Parar.V bola.N:fp
Stop.V ball.N:fp
To stop balls
'to pay attention' [COL]

(23) Ponerse trucha
Poner.V:REFL trucha.N:fs
Put.V:REFL trout.N:fs
To be like a trout
'to pay attention' [MEX]

(24) Pelar los ojos
Pelar.V el.DET:mp ojo.N:mp
Peal.V the.DET:mp eye.N:mp
To open the eyes widely
'to pay attention' [CR]
'to be surprised' [CR]

(25) Parar la oreja
Parar.V la.DET:fs oreja.N:fs
Stand.V the.DET:fs ear.N:fs
To point your ear to a point of interest
'to pay attention' [COL, CR, MEX, PE]

(26) Tomar el pelo
Tomar.V el.DET:ms pelo.N:ms

Take.V the.DET:ms hair.N:ms
To pull someone's hair

'make fun of somebody' [COL]
'try to fool someone' [CR,PE]

(27) Ver la cara
Ver.V la.DET:fs cara.N:fs
See.V the.DET:fs face.N:fs
To see someone's face

'try to fool someone' [COL, CR, MEX, PE]

(28) Comer callado
Comer.V callado.A:ms
Eat.V silent.A:ms
To eat in silence

'don't tell anybody' [COL]

(29) Se le fue la mano
Se.PRON:3s le.PRON:od fue.V:J3s la.DET:fs mano.N:fs
It .PRON:3s to.PRON:od went.V:J3s the.DET:fs hand.N:fs
To have your hand reaching far

'overdid something' [COL, CR, MEX, PE]

(30) Se le pasó la mano
Se.PRON:3s le.PRON:od pasar.V:J3s la.DET:fs mano.N:fs
It .PRON:3s to.PRON:od pass.V:J3s the.DET:fs hand.N:fs
To have your hand over something

'overdid something' [CR, MEX, PE]

(31) Echar los perros
Echar.V el.DET:mp perro.N:mp
Throw.V the.DET:mp dog.N:mp
To throw dogs to someone

'to flirt' [COL, CR, MEX]

(32) Atravesar el caballo
Atravesar.V el.DET:ms caballo.N:ms
Traverse.V the.DET:ms horse.N:ms
To put a horse in the way

'to interrupt' [CR]

(33) Se puso espesa
Se.PRON:3s puso.V:J3s espesa.A:s
It.PRON:3s became.V:J3s thick.A:s
Something becomes thick
'a situation became complicated' [CR]

(34) Meter la pata
Meter.V la.DET:f pata.N:f
Insert.V the.DET:f leg.N:f
To insert the leg somewhere
'screw up' [COL, CR, PE]

(35) Tirar la toalla
Tirar.V la.DET:fs toalla.N:fs
Throw.V the.DET:fs towel.N:fs
To throw a towel to the ground
'to give up' [COL, CR, MEX, PE]

(36) Apretarse el cinturón
Apretar.V:REFL el.DET:ms cinturón.N:ms
Tighten.V:REFL the.DET:ms belt.N:ms
To tighten the belt on the pants
'spend less to save money' [COL, MEX]

(37) Amarrarse los pantalones
Amarrar.V:REFL el.DET:mp pantalón.N:mp
Tighten.V:REFL the.DET:mp pant.N:mp
To tighten up the pants
'impose authority' [CR]

(38) Ponerse los pantalones
Poner.V:REFL el.DET:mp pantalón.N:mp
Put.V:REFL the.DET:mp pant.N:mp
To put on pants
'impose authority' [COL, MEX]

(39) Bailar con la más fea
Bailar.V con.PREP la.DET:fs más.ADV fea.A:fs
Dance.V with.PREP the.DET:fs most.ADV ugly.A:fs
To dance with an ugly lady
'to do the worst part of a job' [CR, MEX]

(40) Se le zafó un tornillo/tuerca
Se.PRON:3s le.PRON:od zafar.V un.DET:ms tornillo.N:ms

It.PRON:3s to.PRON:od loose.V a.DET:ms screw.N:ms
To have a screw loose

'seem crazy' [COL, CR, MEX, PE]

(41) Romper la mano
Romper.V la.DET:fs mano.N:fs
Break.V the.DET:fs hand.N:fs
To break the hand

'to bribe' [PER]

(42) Dar una mordida
Dar.V una.DET:fs mordida.N:fs
Give.V a.DET:fs bite.N:fs
To give someone a bite of something

'to bribe' [COL, CR, MEX]

(43) Lavarse las manos
Lavar.V:REFL la.DET:fp mano.N:fp
Wash.V:REFL the.DET:fp hand.N:fp
To wash your hands

'deny one's responsibility' [COL, CR, MEX, PE]

(44) Sacarse el clavo
Sacar.V:REFL el.DET:ms clavo.N:ms
Remove.V:REFL the.DET:ms nail.N:ms
To take out a buried nail

'to get even with someone' [COL, CR]
'to clear out suspicions' [PE]

(45) Tomar la posta
Tomar.V la.DET:fs posta.N:fs
Take.V the.DET:fs post.N:fs
To grab the relay stick

'to take the lead' [PE]

(46) Tomar la batuta
Tomar.V la.DET:fs batuta.N:fs
Take.V the.DET:fs baton.N:fs
To grab the baton

'to take the lead' [COL,CR]

(47) Estar detrás del palo
Estar.V detrás.PREP del.PREPDET:ms palo.N:ms

Be.V behind.PREP the.PREPDET:ms tree.N:ms
To be standing behind a tree

'When you don't understand or know something that other people is talking about' [CR]

(48) sacar adelante
Sacar.V adelante.ADV
Take-outV. ahead.ADV
Take something out and move forward

'To achieve something' [COL, CR, MEX]

(49) Quedarse corto
Quedar.V:REFL corto.A:s
Stay.V:REFL short.A:s
To be short

'Not having enough money' [COL]
'Not reaching a goal (in any sense)' [COL, CR, MEX]

(50) Sin pelos en la lengua
Sin.PREP pelo.N:mp en.PREP la.DET:fs lengua.N:fs
Without.PREP hair.N:mp in.PREP the.DET:fs tongue.N:fs
Not having hairs in your tongue

'To speak frankly' [COL,CR, MEX]

(51) Se armó la gorda
Se.PRON:3s armar.V la.DET:fs gorda.N:fs
It.PRON:3s assemble.V the.DET:fs fat.N:fs
A fat woman was assembled

'A big problem started' [COL,CR, MEX]

(52) Estar para el gato
Estar.V para.PREP el.DET:ms gato.N:ms
Be.V for.PREP the.DET:ms cat.N:ms
To be like a cat

'Something poorly executed' [PE]

(53) Estar para llorar
Estar.V para.PREP llorar.V:W
Be.V for.PREP cry.V:W
To be about to cry

'Something poorly executed' [CR, MEX, PE]

(54) Estar de llorar
Estar.V de.PREP llorar.V:W
Be.V from.PREP cry.V:W
To be about to cry

'Something poorly executed' [COL]

(55) Estar para el tigre
Estar.V para.PREP el.DET:ms tiger.N:ms
Be.V for.PREP the.DET:ms tiger.N:ms
To be left for a tiger

'To be in a bad state or damaged' [CR]

(56) Echar al agua
Echar.V al.PREPDET:ms agua.N:ms
Throw.V to.PREPDET:ms water.N:ms
To throw someone to a space filled with water

'To be in a bad state or damaged' [CR]

'To push out someone' [COL]

(57) Se le metió el agua
Se.PRON:3s le.PRON:od meter.V el.DET:ms agua.N:ms
It.PRON:3s to.PRON:od introduce.V the.DET:ms water.N:ms
To have water get into you

'To act strangely out of a sudden' [CR]

(58) Aventar la madre
Aventar.V la.DET:fs madre.N:ms
Throw.V the.DET:fs mother.N:fs
To throw a mother to someone

'To insult' [CR, PE]

(59) Pura vida
Pura. vida.N:fs
Pure. life.N:fs
Sheer life

'Very well, cool, hello, goodbye' [CR]

(60) Caer bien
Caer.V bien.A:s
Fall.V good.A:s
To fall in graciously

'To get along well' [CR, MEX]

B Available resources

All the documents and resources that were produced on this project are available online under a Google Drive account created specifically for this purpose. The credentials of this account are:

username: *bi.seminar.ufrt@gmail.com*
password: *biseminar*

Additionally we list here the specific links related to each of the resources:

Spanish MWE database resources

- **XML schema**:
  `https://goo.gl/0x9IY8`

- **XML database**:
  `https://goo.gl/J2moeB`

- **List of 250 Spanish MWEs**:
  `https://goo.gl/pmyXZv`

Corpus for MWE extraction

- **Web form to collect examples**:
  `https://goo.gl/cqotrg`
- **MWE dictionary**: summarised list with the MWE, the associated dialect(s) and the Spanish and English idiomatic meaning
  `https://goo.gl/YHokMy`
- **Corpus**: web link that redirects to the Corpus directory
  `https://goo.gl/ApQGUe`
- **Tutorial**: explains how to register the MWEs examples found in a given corpus, in the web form
  `https://goo.gl/kMgF7v`
- **Video Tutorial**:
  `http://164.15.78.25:81/mwe_example_extraction/`