Towards Lexical Encoding of Multi-Word Expressions in Spanish Dialects

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, Agata Savary
Université François Rabelais Tours, France
agata.savary@univ-tours.fr

Abstract

This paper describes a pilot study in lexical encoding of multi-word expressions (MWEs) in 4 Latin American dialects of Spanish: Costa Rican, Colombian, Mexican and Peruvian. We describe the variability of MWE usage across dialects. We adapt an existing data model to a dialect-aware encoding, so as to represent dialect-related specificities, while avoiding redundancy of the data common for all dialects. A dozen of linguistic properties of MWEs can be expressed in this model, both on the level of a whole MWE and of its individual components. We describe the resulting lexical resource containing several dozens of MWEs in four dialects and we propose a method for constructing a web corpus as a support for crowdsourcing examples of MWE occurrences. The resource is available under an open license and paves the way towards a large-scale dialect-aware language resource construction, which should prove useful in both traditional and novel NLP applications.

Keywords: multi-word expressions, lexical encoding, Spanish dialects

1. Introduction

When Natural Language Processing addresses languages such as Spanish, with a very substantial number of native speakers across various countries, the universalism of the corresponding resources and tools is an important challenge due to dialects. Dialects of the same language can vary e.g. in terms of lexicon, inflection, grammar, or meaning, thus, generic NLP solutions may require customization in order to best fit a given dialect. This problem is similar to the one of automatic processing of closely related languages (Vertan et al., 2013), where one of the challenges it to benefit from the modeling and encoding effort performed for one language in order to deal with another one.

Multi-word expressions (MWEs) such as *sin pelos en la lengua* lit. (to speak) with no hair in the tongue 'frankly', *colgar los zapatos*, lit. to hang shoes 'to die', or *se armó la gorda* lit. a fat woman was assembled 'a big problem started', belong to the most intriguing and challenging dialect specificities. A multi-word expression is understood here as a syntactic structure in which at least two component words (usually including the head-word) are lexicalized, i.e. always realized by the same lexeme, and which shows some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is deemed general grammar rules of the language. In this study we are mostly interested in semantically idiosyncratic expressions, i.e. those whose meaning cannot be compositionally calculated from the meanings of individual components, which is often due to a use of metaphor. We specifically focus on the variability of MWE meaning and use in different dialects of the same language, here 4 Latin American dialects of Spanish.

2. MWEs in Spanish dialects

While lexical encoding and automatic processing of MWEs is an increasingly studied topic (Savary et al., 2015), relatively few attention has been payed so far to the variety of MWEs among different dialects of the same language.[1] This issue proves pervasive in Latin America, notably in Colombia [CO], Costa Rica [CR], Mexico [MEX] and Perú [PE], where the local Spanish dialects, addressed in this work, largely differ in their use of MWEs.

For example, *estar limpio* (literally 'to be clean') is a MWE meaning 'to be out of money' in Spanish from Costa Rica [CR] but not in the 3 other dialects. The same meaning can be represented though by a lexically different MWE in another dialect, e.g. *estar pelado* [CO] (lit. 'to be naked') 'to be out of money'. One MWE can be valid in several dialects and have the same meaning – e.g. *echar los perros* [CO,CR,MEX] (lit. 'to throw the dogs') 'to flirt' – or different meanings – e.g. *hablar paja* (lit. 'to talk straw') 'to tell lies' [CR], 'to make small talk' [CO], or 'to speak wonderfully' [MEX]. Finally, some MWEs have many possible meanings, even within a single dialect. For example, the phrase *pura vida* (lit. 'pure life') is used in Costa Rica to say hello, say goodbye, describe a person who is very nice and easy going, express how one is 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 exciting, and so on.

This work is a pilot study for representing Spanish MWEs in a computational lexicon taking dialect specificities into account, as well as for constructing a dialect-specific corpus of MWE examples. A more detailed description of these contributions can be found in (Arauco et al., 2015).

3. Data model

The proposed data model adapts and extends the one proposed in (Itai and Wintner, 2008) so as to: (i) represent dialects in which a given MWE is valid together with its (dialect-dependent) meanings, (ii) link MWEs that are different in form but same in meaning in one or more Spanish dialects. A dozen of linguistic properties are taken into account, including the following:

- meaning (identifier of the English translation of the MWE),
- dialect (CO, CR, MEX, PE),

[1] See, notably (Hawwari et al., 2014), who put forward a framework for a computational lexicon of MWEs in Egyptian Arabic as opposed to Modern Standard Arabic.

- language register (colloquial, casual etc.),
- passivization (can the MWE be expressed in the passive voice while keeping its idiomatic reading?),
- partial inflection (can a component of the MWE inflect?),
- inflection degree (is the component non-variable, partly inflectionally flexible or can it take any of its inflected forms in the MWE?),
- modification (can components be modified or substituted by external words?)

Most of these properties can apply to the entire MWE, at a component level or both, and be either dialect-specific or generic. The challenge is to represent dialect-related specificities while avoiding redundancy of the data common for all dialects.

Appendix A shows a sample encoding of the MWE *aventar la madre* [CR,PE] (lit. 'throw the mother') 'to insult', which consists of three components identified morphologically and glossed in the ⟨baseTokensList⟩ element. Only the first of them allows inflections. The ⟨properties⟩ mentioned at the level of the whole MWE include: (i) allowing no additions of external elements, (ii) allowing substitutions by the ⟨substituteToken⟩ *mentar* 'mention', (iii) allowing inflection in at least one component[2], (iv) belonging to the vulgar ⟨languageRegister⟩, (v) allowing ⟨passivization⟩. The ⟨paths⟩ specify how different morpho-syntactic variants of this MWE can be constructed with different inflected forms of its ⟨baseToken⟩s and ⟨substituteToken⟩s. The first ⟨path⟩ describes the canonical form *aventar la madre*, while the second one represents the variant where the ⟨substituteToken⟩ *mentar* replaces *aventar*. Since no @dialect attribute is mentioned at the level of any of these properties, they all apply to the dialects mentioned in the ⟨meaningInDialect⟩ elements (here: CR and PE).

Appendix B shows an extract of another example where the MWE *hablar paja* [CO,CR,PE] (it. 'to talk straw') has a different meaning in each of the 3 dialects (represented in Appendix C). It ⟨allowsAdditions⟩ in all 3 dialects but the allowed ⟨additionalToken⟩s are different in Peru than in Costa Rica and Colombia, as indicated by the different values of the @dialect attribute. Namely, in the Costa Rican and Colombian dialects the additional token is either the quantifier *mucha* 'a lot' modifying the nominal complement *paja* 'straw', or the adverbial *solo* 'only' modifying the same nominal complement (if occurring at position 2: *hablar solo paja* 'talk only straw') or the whole verbal phrase (if at position 1: *solo hablar paja* 'only talk straw') . The Peruvian dialect, in turn, only allows for the adverb *muy* 'very' modifying *paja*, which suggests that this complement shifts towards an adjectival or adverbial reading in this MWE.

[2]Thus, a functional dependency needs to be defined between the ⟨allowsInflections⟩ element in ⟨properties⟩ and the @allowsInflections attribute in ⟨baseToken⟩.

4. Lexical resource and corpus

The lexical resource of sample MWEs has been created. We have gathered about 250, mainly verbal, MWE examples, 40 of which have been described by native speakers of the four Spanish dialects. The XML database containing the 40 MWE descriptions, the associated XML schema, and the 250 MWE examples with literal and idiomatic translations, are available[3] under the terms of the 2-clause BSD license. In parallel, we wish to assign three types of corpus occurrences to these MWEs (see Appendix D):

- positive examples, where a given MWE occurs with its idiomatic meaning,
- neutral examples, where it has a literal (compositional) meaning,
- negative examples, where all lexicalized components of the MWE occur but their syntactic dependencies are not those assumed by the MWE.

We also started collecting images to illustrate these types of occurrences (see Appendix E).

An automated corpus construction method has been developed to collect positive, neutral and negative corpus occurrences. It is based on query expansion. A given MWE is first tokenized, its stop words are identified, and the remaining components are processed by the AGME morphological tool[4] using the FreeLing[5] database with 26,000 Spanish lemmas. Each component is expanded to lists of all its inflected forms. A random selection of around 100 combinations of these inflected forms (one per component) is then submitted to the BootCat[6] web crawler (one form combination per query), which is parameterized with URLs specific to particular Spanish dialects, as well with the NEAR value, which requires the searched words to occur within a window of a restricted length. As an output of the process the web crawler creates a text file (one per query) containing the appended contents from all the web pages where the data were found. These text corpora are then stored and linked to a crowdsourcing procedure driven by a web form, where speakers of different dialects (recruited notably via social networks) can search for positive, neutral and negative examples and store them together with their source URLs. They are also encouraged to search for images illustrating these examples. Corpus examples for 25 MWEs and a couple of pictures have been registered so far.

5. Applications

As the size and coverage of this lexicon grow, it can become a useful resource for various kinds of MWE-aware NLP applications such as parsers, MWE extractors, MWE lemmatizers, word-sense disambiguation tools, etc. Moreover, its dialect-oriented features can prove useful in more specific applications such as:

[3]http://www.info.univ-tours.fr/~savary/English/resourcesASgb.html#Spanish-WMEs

[4]http://www.cic.ipn.mx/~sidorov/agme/

[5]http://nlp.lsi.upc.edu/freeling/

[6]http://bootcat.sslmit.unibo.it/

- customization of generic Spanish NLP tools for particular dialects,
- MWE-aware machine translation or computer-aided translation of texts among various Spanish dialects,
- dialect identification (especially if the MWE entries are enriched with probabilities of their idiomatic vs. literal readings),
- automatic MWE identification and disambiguation (based on positive, neutral and negative examples),
- second language/dialect learning (especially due to images associated with the three types of examples).

6. Limitations and future work

The web-based collaborative procedure designed to search for corpus occurrences of MWEs proved difficult for highly idiomatic examples. Namely, the majority of web documents retrieved for such MWEs by this procedure are pages defining these very idioms in on-line dictionaries and thesauruses. The sample sentences contained therein are often artificially constructed and, thus, do not fulfill the usual criteria of using the web as a corpus.

The lexical encoding schema proposed in this work accounts for morphological and partly for syntactic constraints imposed on MWEs. Many MWEs, however, especially verbal ones, enter into other complex syntactic relationships with the remaining words in a sentence (agreement, control, word order, etc.). For instance, the subject of a sentence containing the idiom *hacerse el loco* (lit. 'to make oneself the crazy') 'to pretend not to understand' must agree in person and number with the head verb *hacer* 'make' and in number and gender with the nominalized adjectival predicate *el loco* 'the crazy', e.g. *ella se hizo la loca* 'she pretends not to understand'. The same holds for the reflexive marker *se* 'self', which detaches from the infinitive and is placed before the verb. Such phenomena can only be fully accounted for within a (deep syntactic) grammar of a language. Note also that if such a lexicon is made to cooperate with a grammar then some morphological data might prove redundant between the two resources (e.g. the data on agreement on the internal components).

Some other formalisms, such as DUELME (Grégoire, 2010), suggest a two-level encoding, in which the lexicon only encodes those properties which are not accounted for in the grammar. MWE *equivalence classes* group MWEs with common syntactic behavior, while MWE *entries*, assigned to equivalence classes, represent the entry-specific idiosyncrasies. This proposal, however, still does not allow to fully avoid data redundancy, while preserving the precision of the linguistic description: if we wish to keep the number of equivalence classes reasonably low, idiosyncratic properties common to several MWE entries have to be repeated in each of them.

We think that a solution to this problem might come from an *object-oriented* lexical encoding, in which MWE classes would be organized into a hierarchy, so that lower classes would inherit more general properties from higher classes and add some more specific properties. MWE entries would then be objects assigned to subsets of classes. In this way we could achieve a "discrete continuum" between the lexicon and the grammar, which might accurately represent the intuition the very nature of MWEs. Dialect-specific properties should also find a natural representation in such a setting. First steps towards such an object-oriented representation of MWEs within the XMG metagrammar framework (Lichte and Petitjean, 2015) have been taken in (Lichte et al., 2016).

7. Acknowledgements

This work is an outcome of a student project carried out within the Erasmus Mundus Master's program "Information Technologies for Business Intelligence"[7]. It was supported by the IC1207 COST action PARSEME[8]. We are grateful to prof. Shuly Wintner for his valuable insights into lexical encoding of MWEs.

References

Arauco, A., Bogantes, D., Rodríguez, A., Rodríguez, E., and Savary, A. (2015). Representation and Identification of Multiword Expressions in different Spanish Dialects. Technical Report 314, Laboratoire d'informatique, François Rabelais University of Tours, France. http://www.info.univ-tours.fr/~savary/enseignementAS.html#BI-sem-2015.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2).

Hawwari, A., Attia, M., and Diab, M. (2014). A Framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 48–56, Doha, Qatar, October. Association for Computational Linguistics.

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

Lichte, T. and Petitjean, S. (2015). Implementing semantic frames as typed feature structures with XMG. *J. Language Modelling*, 3(1):185–228.

Lichte, T., Parmentier, Y., Petitjean, S., Savary, A., and Waszczuk, J. (2016). Separating the regular from the idiosyncratic: A constraint-based lexical encoding of MWEs using XMG. http://typo.uni-konstanz.de/parseme/index.php/2-general/156-selected-posters-struga-7-8-april-2016.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Butt, M., Losnegaard, G. S., Escartín, C. P., Waszczuk, J., Constant, M., Osenova, P., and Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *Language & Technology Conference (LTC'15)*, Poznań, Poland.

[7] http://it4bi.univ-tours.fr/

[8] http://www.parseme.eu

Cristina Vertan, et al., editors. (2013). *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, September.

Appendix A

Encoding of a sample MWE *aventar la madre* (lit. 'throw the mother') 'to insult'

```
- <mwe id="MWE10" mweText="aventar la madre" length="3">
    <meaningInDialect id="MWE10ISCR" meaning="IS" dialect="CR"/>
    <meaningInDialect 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>throw</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"/>
        <node token="MWE10_BTkn3" fixedAnalysis="N.fs"/>
      </path>
    - <path>
        <node token="MWE10_ATkn1"/>
        <node token="MWE10_BTkn2" fixedAnalysis="DET.fs"/>
        <node token="MWE10_BTkn3" fixedAnalysis="N.fs"/>
      </path>
    </paths>
  - <sourcesInCorpus>
    - <source typeOfExample="possitive">
        <sourceName>Taringa</sourceName>
      - <link>
          http://www.taringa.net/comunidades/taringamexico/7301812/Y-Que-Listos-Para-Aventarle-La-Madre-a-EPN.html
        </link>
      </source>
    + <source typeOfExample="neutral"></source>
    + <source typeOfExample="negative"></source>
    </sourcesInCorpus>
  </mwe>
```

Appendix B

Extract of the encoding of a sample MWE *hablar paja* (lit. 'to talk straw') with 3 different meanings according to the dialect.

```
- <mwe id="MWE27" mweText="hablar paja" length="2">
    <meaningInDialect id="MWE27TICR" meaning="TI" dialect="CR"/>
    <meaningInDialect id="MWE27STCO" meaning="ST" dialect="CO"/>
    <meaningInDialect id="MWE27SWPE" meaning="SW" dialect="PE"/>
  - <properties>
      <allowsAdditions value="true"/>
      <allowsSubstitutions value="true" dialects="PE"/>
      <allowsInflections value="true"/>
      <languageRegister value="Colloquial"/>
      <passivization value="false"/>
    - <additionalTokensList>
      - <additionalToken id="MWE27_ATkn1" isStopWord="true" possition="2" allowsInflections="false"
        allowsSubstitutions="true" analysis="A.fs" dialects="CR CO">
          <wordES>mucha</wordES>
          <wordEN>a lot</wordEN>
        </additionalToken>
      - <additionalToken id="MWE27_ATkn2" isStopWord="true" possition="2" allowsInflections="false"
        allowsSubstitutions="true" analysis="ADV" dialects="CR CO">
          <wordES>solo</wordES>
          <wordEN>only</wordEN>
        </additionalToken>
      - <additionalToken id="MWE27_ATkn3" isStopWord="true" possition="1" allowsInflections="false"
        allowsSubstitutions="true" analysis="ADV" dialects="CR CO">
          <wordES>solo</wordES>
          <wordEN>only</wordEN>
        </additionalToken>
      - <additionalToken id="MWE27_ATkn4" isStopWord="true" possition="2" allowsInflections="false"
        allowsSubstitutions="true" analysis="ADV" dialects="PE">
          <wordES>muy</wordES>
          <wordEN>very</wordEN>
        </additionalToken>
      </additionalTokensList>
      <!-- other properties -->
    </properties>
  </mwe>
```

Appendix C

Extract of the encoding of sample MWE meanings, dialects and inflectional features.

```
- <meaninigs>
    <meaning id="IS" meaning="To insult"/>
    <meaning id="TI" meaning="To tell lies"/>
    <meaning id="ST" meaning="To have a small talk"/>
    <meaning id="SW" meaning="To speak wonderfully"/>
    <!-- more meanings -->
  </meaninigs>
- <dialects>
    <dialect id="CO" country="Colombia"/>
    <dialect id="CR" country="Costa Rica"/>
    <dialect id="MEX" country="Mexico"/>
    <dialect id="PE" country="Peru"/>
  </dialects>
- <inflections>
    <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"/>
    <!-- more inflections -->
  </inflections>
```

Appendix D

Positive (1), neutral (2) and negative (3) example of corpus occurrence for the MWE *colgar los zapatos* (lit. to hang shoes) 'to die'

(1) Pienso gastarme hasta el último peso, espero antes de **colgar los zapatos**. [MEX]
(lit.) I plan to spend until my last penny, hopefully before hanging the shoes.
'I plan to spend until my last penny, hopefully before I die.'

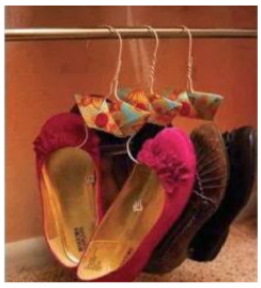
(2) Una idea interesante es modificar una percha de alambre para **colgar los zapatos** y, de esta manera, ahorrar espacio.
'An interesting idea is to change a wire hanger to hang the shoes and, thus, save space.'

(3) Los famosísimos **zapatos** de la serie Sexo en Nueva York ahora están disponibles para comprarlos y **colgártelos**!
'The famous shoes of the Sex in New York TV show are now available for purchase and wearing!'

Appendix E

Images illustrating the positive, neutral and negative occurence of the MWE from Appendix D.

Positive



Neutral

Negative