Towards Lexical Encoding of Multiword Expressions in Spanish Dialects [WG1]

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

1 MWEs in Spanish dialects

While lexical encoding and automatic processing of Multi-Word Expressions (MWEs) is an increasingly studied topic, 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. *ponerse las pilas* (lit. 'put the batteries') 'to start doing something seriously' [CO], 'to do things in a better way' [CR], 'to be more active' [MEX], or 'to do things faster' [PE]. 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).

2 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 – e.g. meaning, dialect, language register, passivization, partial inflection, etc. – most of which 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.

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

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 both dialects mentioned in the ⟨meaningInDialect⟩ elements.

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 ⟨allows Additions⟩ 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.

3 Lexical resource and corpus

The lexical resource of sample MWEs is currently being created. We have gathered about 250, mainly verbal, MWE examples, 40 of which have been described by native speakers of the four Spanish dialects.[3] In parallel, we wish to assign three types of corpus occurrences to these MWEs: (i) positive examples, where a given MWE occurs with its idiomatic meaning, (ii) neutral examples, where it has a literal (compositional) meaning, (iii) negative examples, where all lexicalized components of the MWE occur but their syntactic dependencies are not those assumed by the MWE (see Appendix D). We also started collecting images to illustrate these types of occurences (see Appendix E). The automated construction method 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 analyzer[4] using the FreeLing[5] database with 26,000 Spanish lemmas. The lemmatized components are expanded to lists of all their 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

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

[3]The XML database containing the 40 MWE descriptions, as well as the associated XML schema, are available under the terms of the 2-clause BSD license at http://www.info.univ-tours.fr/~savary/English/resourcesASgb.html#Spanish-WMEs. A more complete set of corpus examples should be available shortly.

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

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

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

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 can search for positive, neutral and negative examples and store them together with their source URLs. Corpus examples for 25 MWEs have been registered so far.

4 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]. We are grateful to prof. Shuly Wintner for valuable insights into lexical encoding of MWEs.

References

[Arauco et al.2015] Alejandro Arauco, Diana Bogantes, Alejandro Rodríguez, Eric Rodríguez, and Agata Savary. 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.

[Hawwari et al.2014] Abdelati Hawwari, Mohammed Attia, and Mona Diab. 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 and Wintner2008] Alon Itai and Shuly Wintner. 2008. Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

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

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 series 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