Towards Lexical Encoding of MWEs in Spanish dialects [WG1]

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, Agata Savary
Université François-Rabelais Tours, Blois, France

Challenges with Spanish MWEs

- Inflections (on gender, number, tense, mood, ...):
  - *hacerse el loco* [to act crazy] → pretend not to see something
  - *me hice el loco*, *se hace la loca*, *se hicieron los locos*, *me hago la loca*
- Possibility of adding new words:
  - *hablar paja* [have straw out of your mouth] → hablar paja
  - *hablar* **pura** *paja*
- Possibility of substituting words:
  - *tomar la posta* [to grab the relay stick] → to take the lead
  - *tomar la* **batuta**
- Possibility of omitting base words:
  - *se armó la gorda* [a fat woman was assembled] → a big problem started
  - *se armó*

More Spanish dialects = more challenges

- Unique expressions
  - *estar limpio*$_{\text{CR}}$ [to be clean] → to be out of money
- Same meaning, different expression
  - *estar aguja*$_{\text{COL}}$ [to be needle] → to be out of money
- Same expression, same meaning
  - *echar los perros*$_{\text{COL,CR,MEX}}$ [to throw the dogs] → to flirt
- Same expression, different meaning
  - *ponerse las pilas* [to put on the batteries]
    - → to start doing sth seriously$_{\text{COL}}$
    - → to do things in a better way$_{\text{CR}}$
    - → to be more active$_{\text{MEX}}$
    - → to do things faster$_{\text{PE}}$

Database example

- *aventar la madre* [to throw the mother] → to insult$_{\text{CR,PE}}$

```
- <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>
```

- the XML schema extends [1]
- challenges:
  - represent dialect-related specificities while avoiding redundancy
  - link dialect-specific MWEs with the same meaning

Database example with variations in meaning and in allowed insertions

- *hablar paja* [to talk straw]

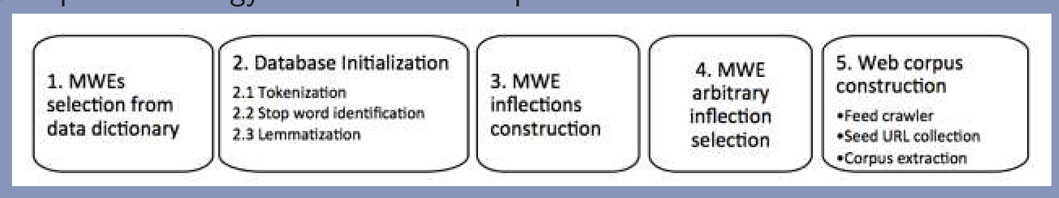
```
- <mwe id="MWE27" mweText="hablar paja" length="2">
    <meaningInDialect id="MWE27TICR" meaning="TI" dialect="CR"/>
    <meaningInDialect id="MWE27STCO" meaning="ST" dialect="CO"/>
    <meaningInDialect id="MWE27SWPE" meaning="SW" dialect="PE"/>
  - <properties>
      <allowsAdditions value="true"/>
      <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>
```

- 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>
```

Lexical Resource and Corpus

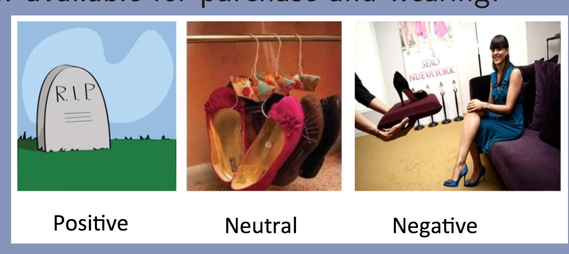
- 250 MWE examples were gathered, natives speakers of the four Spanish dialects have described 40 of them.
- 5-step methodology to construct a corpus based on the web:



MWEs examples from corpus and support images

Extraction of 3 types of MWE examples from the corpus by crowdsourcing:

- **Positive example:** a given MWE occurs with its idiomatic meaning
  - *Pienso gastarme hasta el último peso, espero antes de colgar los zapatos.*$_{\text{MEX}}$ [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.
- **Neutral example:** a given MWE has a literal (compositional) meaning.
  - *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.
- **Negative example:** all lexicalized components of the MWE occur but their syntactic dependencies are not those assumed by the MWE
  - *Los famosísimos zapatos de la serie Sexo en Nueva York ahora están disponibles para comprarlos y colgártelos!*$_{\text{MEX}}$ → The famous shoes of the Sex in New York TV show are now available for purchase and wearing!



Bibliography

[1] Alon Itai and Shuly Wintner.
Language resources for Hebrew.
Language Resources and Evaluation, 42(1):75–98, 2008.