Putting Figures on Influences on Moroccan Darija from Arabic, French and Spanish using the WordNet

Khalil Mrini
Ecole Polytechnique Fédérale de Lausanne
khalil.mrini@epfl.ch

Francis Bond
Nanyang Technological University
bond@ieee.org

Introduction

Context

Moroccan Darija, local Arabic variant, is spoken by 90.9% of population

Major influences:
- Arabic: official, Afro-Asiatic
- Tamazight: official, Afro-Asiatic

Minor, Colonial influences:
- French: widespread, Indo-European
- Spanish: limited, Indo-European

Objective: automatic estimation of linguistic influences on Moroccan Darija

Data: Open Multilingual WordNet (OMW) (Bond and Foster, 2013), which contains the Moroccan Darija (MDW) (Mrini and Bond, 2018), Arabic (Black et al., 2016; Abouennour et al., 2013), French (Sagot and Fler, 2008), Spanish (Gonzalez-Agirre et al., 2012) wordnets

Results and Discussion

• Comparisons based on automatically linked synsets of the MDW and manually validated ones.

Comparison with automatically linked synsets

<table>
<thead>
<tr>
<th>Comparison with:</th>
<th>Arabic</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of links to Moroccan synsets</td>
<td>7,958</td>
<td>11,605</td>
<td>10,167</td>
</tr>
<tr>
<td>– excluding synsets with only multi-word expressions</td>
<td>6,702</td>
<td>9,954</td>
<td>8,612</td>
</tr>
<tr>
<td>Average normalised Levenshtein distance</td>
<td>0.4619</td>
<td>0.7337</td>
<td>0.7521</td>
</tr>
<tr>
<td>Number of synsets with one or more word pairs at least 60% similar</td>
<td>2,816</td>
<td>278</td>
<td>188</td>
</tr>
<tr>
<td>Percentage of synsets with one or more word pairs at least 60% similar</td>
<td>42.02%</td>
<td>2.79%</td>
<td>2.18%</td>
</tr>
</tbody>
</table>

• If confidence scores are used as weights: average normalised Levenshtein distance of Moroccan Darija is 0.4701 with Arabic, 0.7598 with French, 0.7775 with Spanish.

• To diminish randomness of similarity, a threshold is established empirically at 60%.

• Arabic and Moroccan Darija: Closest pair of languages in similarity are Portuguese and Galician (average Levenshtein distance of 0.4760), two independent languages.

• Moroccan Darija with French and Spanish: Out of the synsets more than 60% similar, 95 are common. Future work to allow distinction of origin of influence.

• Moroccan lemmas of unknown origin: 2,736. They include:
  - Lemmas originating from Arabic (nzel, to go down), Spanish (serbisa, beer), not detected due to errors in linking.
  - Sizeable proportion of Tamazight origin (seqsi, to ask). Influence particularly visible on words starting with fa- and ending in -t (bazeligt, tabennaguet, etc.).

Comparison with manually validated synsets

<table>
<thead>
<tr>
<th>Comparison with</th>
<th>Average distance</th>
<th>At least 60% similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>French</td>
<td>Spanish</td>
</tr>
<tr>
<td>The 12,224 synsets that form the total</td>
<td>0.4619</td>
<td>0.7337</td>
</tr>
<tr>
<td>The 617 manually validated synsets</td>
<td>0.4393</td>
<td>0.7544</td>
</tr>
</tbody>
</table>

• Differences in figures small enough to prove linking noise was not an issue. However, the number of lemma pairs at least 60% similar is too small to give clear validation.

References


