Contributions

- Revision of Curras[1]:
  - Palestinian morphological corpus
- Creation of Baladi:
  - Newly annotated Lebanese corpus
- A more Levantine corpus [2]:
  - Corpora compatibility, POS normalization and unification, and Lemma coherence
  - https://portal.sina.birzeit.edu/curras

Arabic is low resourced with several dialects

Modern standard Arabic (MSA):
- News, education, media – has more resources

Classical Arabic:
- – traditional and literature

Dialects: spoken daily, social media
- – region and interlingual influence

Dialect families:
- Moroccan, Egyptian, Sudanese, Levantine, Iraqi and Khali (Gulf)

CODA annotation guidelines

- The correct and unified spelling of the token
- Followed Curras CODA

<table>
<thead>
<tr>
<th>Token</th>
<th>CODA</th>
</tr>
</thead>
<tbody>
<tr>
<td>bâlîn</td>
<td>bâlîn → bâlîn</td>
</tr>
<tr>
<td>bâlîhn</td>
<td>bâlîhn → bâlîhn</td>
</tr>
<tr>
<td>târy</td>
<td>târy → târy</td>
</tr>
<tr>
<td>hîdy</td>
<td>hîdy → hîdy</td>
</tr>
<tr>
<td>ñûmmn</td>
<td>ñûmmn → ñûmmn</td>
</tr>
</tbody>
</table>

Prefix annotation guidelines

- Based on SAMA[3] prefix tagset

<table>
<thead>
<tr>
<th>Prefix(es)</th>
<th>Surface Tag</th>
<th>Feature</th>
<th>Stem Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>ñûmmn</td>
<td>ñûmmn</td>
<td>-</td>
<td>ñûmmn</td>
</tr>
<tr>
<td>ñûmmn</td>
<td>ñûmmn</td>
<td>-</td>
<td>ñûmmn</td>
</tr>
<tr>
<td>ñûmmn</td>
<td>ñûmmn</td>
<td>-</td>
<td>ñûmmn</td>
</tr>
<tr>
<td>ñûmmn</td>
<td>ñûmmn</td>
<td>-</td>
<td>ñûmmn</td>
</tr>
</tbody>
</table>

Stem annotation guidelines

- <Stem /POS> schema: Used Curras stems

POS, Person, Aspect, Gender, Number
- Same SAMA tags sets

Curras revisions

- Verified Person, Aspect, Gender, and Number features
- Revised almost all annotations
- Generated a minimum set of unique morphological solutions
- Used it to annotate Baladi

Tokenization and POS

- No parsing errors
- Backwater transliterations
- CODA corresponds to prefixes-stem-suffixes
- Prefix/suffix(es) in the predefined sets
- Stem/POS in the SAMA POS tagset.

Lemma annotation guidelines

- MSA Lemma, Dialect Lemma
- SAMA lemmas: If none, add one in the same way, e.g., بَلَّأ: بَلَّأ

Suffix functiona words

- Verbal, Noun, Adjective
- Derived words
- Manually revised MSA and DA lemmas
- MSA lemmas are linked with SAMA lemmas,
- New lemmas are marked with “_” 0
- DA lemmas are linked with MSA lemmas (if different)

References