The AMARA Corpus: Building Parallel Language Resources for the Educational Domain
Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, Stephan Vogel

**Motivation**

Massive Online Open Courses (MOOCs) bridge financial/geographical gap

Thousands of lectures are available (70% English)

**Source**

![Pie chart showing distribution of sources]

Amara

Is a web-based platform for creating, editing and managing video subtitles

Is driven by volunteers

Hosts many educational videos subtitled/translated

**Building Parallel Resources for Education**

Crawl

(a) download subtitles

- 43K videos
- 160 language
- 121K translations/transcripts

(b) keep educational videos

Validate

Content of the subtitles

(a) completeness:
- Discard empty or incomplete

(b) correct language:
- Using Cybozu language detector

- 12.2K videos
- 20 language
- 34K translation/transcript

Align

Parallel segments

(a) strict synchronization:
- Accept only parallel segments w/identical ds and timestamps

(b) cascaded synchronization:
- Re-align the discarded segments using length statistics and information from a bilingual dictionary

**Sample Translations**

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>When the arm will return to its original position</td>
<td>لطرح 3 من كل الطرفين</td>
</tr>
<tr>
<td>Then the arm will return to its original position</td>
<td>ما هو الرمز البياني لـ</td>
</tr>
<tr>
<td>Entonces esto es igual a 4x al cuadrado menos 2x más 8.</td>
<td>Sopra es igual a 4x squared minus 2x plus 8.</td>
</tr>
</tbody>
</table>

**Acknowledgements**

We thank the Amara.org staff, especially Nicholas Reville, Aleli Alcala and Dean Jansen for their support in the development of this corpus.