Languages with no training data

<table>
<thead>
<tr>
<th>Language</th>
<th><strong>Naija</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. apply English tokenizer</td>
<td></td>
</tr>
<tr>
<td>2. &quot;translate&quot; words to English</td>
<td></td>
</tr>
<tr>
<td>3. apply English tagger and parser</td>
<td></td>
</tr>
<tr>
<td>4. copy form to lemma, remove final -s</td>
<td></td>
</tr>
</tbody>
</table>

Wiki:
A bai shu giv mai broda (I bought shoes that I gave to my brother)

Bible:
Jiles Kraist wey dem born for David and Abraham famili, na em stori bi dis.

jw.org:
We come from different different place and we dey speak different different language.

Languages with small training data

<table>
<thead>
<tr>
<th>Language</th>
<th><strong>Armenian</strong></th>
<th><strong>Buryat</strong></th>
<th><strong>Kazakh</strong></th>
<th><strong>Kurmanji</strong></th>
<th><strong>Up. Sorbian</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Armenian 0.57</td>
<td>Buryat 0.45</td>
<td>Kazakh 0.44</td>
<td>Kurmanji 0.52</td>
<td>Up. Sorbian 0.40</td>
</tr>
<tr>
<td></td>
<td>Latvian 0.56</td>
<td>Hindi 0.41</td>
<td>Turkish 0.33</td>
<td>Latin 0.47</td>
<td>Polish 0.60</td>
</tr>
<tr>
<td></td>
<td>Estonian 0.51</td>
<td>Uyghur 0.38</td>
<td>Uyghur 0.29</td>
<td>Greek 0.45</td>
<td>Czech 0.51</td>
</tr>
</tbody>
</table>

Treebank translation
- word-alignment (FastAlign) on OpenSubtitles2018 (Opus)
- each word is translated by the target word that was most frequently aligned to it

Pretrained embeddings
When the UDPipe parser is trained, we use pretrained word-embeddings by Bojanovski et al. (2016) on Wikipedia texts

UniMorph morphology
Universal morphology annotation (Sylak-Glassman, 2018) uses different features form UD, mapping needed:
1. Presn=1
2. Number=Sg
3. Mood=Cnd
4. Case=Acc
5. Aspect=Perf
6. Person=1

Combining multiple parsers
Several different parse trees are combined together.
Each tree may have different weight (expected performance on target treebank)
Maximum spanning tree is used to chose the best result.

Overall results on low-resource languages

<table>
<thead>
<tr>
<th><strong>Source 1</strong></th>
<th><strong>Source 2</strong></th>
<th><strong>Source 3</strong></th>
<th><strong>Target</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>LAS</td>
<td>LAS</td>
<td>LAS</td>
<td>BLEX</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>27.89</td>
<td>6.13</td>
<td>13.98</td>
<td>1</td>
</tr>
</tbody>
</table>

Languages without training data

Preprocessing
- aligned to it
- OpenSubtitles2018 (Opus)

Pre-trained embeddings
- word-alignment (FastAlign) on

UniMorph morphology
- uses

Combining multiple parsers
- several
- expected performance on target treebank

Overall results on low-resource languages
- source 1:
- source 3:
- target:

Languages with small training data

Pretrained embeddings
- when the UDPipe parser is trained, we use pretrained word-embeddings by Bojanovski et al. (2016) on Wikipedia texts

Post-correction:
- if the word is found in UniMorph, change its tag, lemma and features

Combining multiple parsers
- several different parse trees are combined together
- each tree may have different weight (expected performance on target treebank)
- maximum spanning tree is used to chose the best result

Overall results on low-resource languages
- source 1:
- source 3:
- target: