MaiBaam – A multi-dialectal Bavarian Universal Dependency treebank

Verena Blaschke, Barbara Kovačić, Siyao (Logan) Peng, Hinrich Schütze, Barbara Plank

Bavarian

- Spoken by 10M+ people in DE, AT, IT
- No orthography
- 3 main dialect groups (+ transition areas)
- Closely related to German, yet morphosyntactic differences

I mecht wissn, wej lång das'sd nu brauchst.
I mechad gern wissn, wia lang du no brauchst.
I mecht gearn wissn, wia long du no brausch.
DE Ich möchte wissen, wie lange du noch brauchst.
EN I want to know how long you'll still take.

Treebank stats

<table>
<thead>
<tr>
<th>Dialect group</th>
<th>Toks</th>
<th>Genre</th>
<th>Toks</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>833</td>
<td>Wiki</td>
<td>7 988</td>
</tr>
<tr>
<td>North/Central</td>
<td>793</td>
<td>Grammar ex.</td>
<td>2 485</td>
</tr>
<tr>
<td>Central</td>
<td>3 303</td>
<td>Non-fiction</td>
<td>2 019</td>
</tr>
<tr>
<td>South/Central</td>
<td>1 130</td>
<td>Social</td>
<td>1 599</td>
</tr>
<tr>
<td>South</td>
<td>995</td>
<td>Fiction</td>
<td>932</td>
</tr>
<tr>
<td>Underspecified</td>
<td>7 969</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15 023</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data, paper & annotation guidelines

 Parsing baselines

- Best system: UDPipe–GSD
  80.3% POS accuracy, 65.8% LAS
  
  Ample room for improvement! Can you beat it?

- Input representations seem to be an important factor in parsing/tagging performance (full words vs. subwords vs. characters)
- More experiments in paper

Gold (top) vs. prediction by best system (bottom)
“The Lammer (river) has fairly clean water”

This research is supported by the ERC Consolidator Grant DIALECT 101043235.