Do dependency parsing metrics correlate with human judgments?

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, Anders Søgaard

Center for Language Technology, University of Copenhagen, Denmark
Department of Linguistics, University of Zagreb, Croatia
bplank@cst.dk

A systematic comparison between 7 dependency parsing evaluation metrics and human judgments of overall parse quality.
A novel dataset of 984 sentences annotated with human judgments for five languages.

- Human-metric correlation is lower for dependency parsing than for other NLP tasks.
- Inter-annotator agreement is sometimes higher than agreement between judgments and metrics.
- Humans have a preference for attachment over labeling, and attachment closer to the root is more important.

### Datasets

- **English**
- **Spanish**
- **Danish**
- **German**
- **Croatian**

### Metrics

- Unlabeled attachment score (UAS)
- Labeled attachment score (LAS)
- Label accuracy (LA)
- Unlabeled complete predicates (UCP)
- Labeled complete predicates (LCP)
- Neutral Edge Direction (NED) (Schwartz et al., 2011)
- Tree Edit Distance (TED) (Tsarfaty et al. 2011; 2012)

### Results

<table>
<thead>
<tr>
<th>Annotator</th>
<th>English LAS</th>
<th>English UAS</th>
<th>Spanish LAS</th>
<th>Spanish UAS</th>
<th>Danish LAS</th>
<th>Danish UAS</th>
<th>German LAS</th>
<th>German UAS</th>
<th>Croatian LAS</th>
<th>Croatian UAS</th>
<th>All LAS</th>
<th>All UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>0.818</td>
<td>0.848</td>
<td>0.808</td>
<td>0.828</td>
<td>0.745</td>
<td>0.765</td>
<td>0.547</td>
<td>0.478</td>
<td>0.297</td>
<td>0.466</td>
<td>0.540</td>
<td>0.457</td>
</tr>
<tr>
<td>de</td>
<td>0.710</td>
<td>0.690</td>
<td>0.635</td>
<td>0.710</td>
<td>0.575</td>
<td>0.565</td>
<td>0.541</td>
<td>0.437</td>
<td>0.331</td>
<td>0.453</td>
<td>0.397</td>
<td>0.425</td>
</tr>
<tr>
<td>en</td>
<td>0.728</td>
<td>0.715</td>
<td>0.705</td>
<td>0.660</td>
<td>0.658</td>
<td>0.525</td>
<td>0.601</td>
<td>0.663</td>
<td>0.644</td>
<td>0.603</td>
<td>0.581</td>
<td>0.554</td>
</tr>
<tr>
<td>es</td>
<td>0.601</td>
<td>0.663</td>
<td>0.644</td>
<td>0.601</td>
<td>0.652</td>
<td>0.635</td>
<td>0.581</td>
<td>0.581</td>
<td>0.581</td>
<td>0.554</td>
<td>0.581</td>
<td>0.554</td>
</tr>
<tr>
<td>fr</td>
<td>0.800</td>
<td>0.755</td>
<td>0.700</td>
<td>0.730</td>
<td>0.730</td>
<td>0.705</td>
<td>0.570</td>
<td>0.580</td>
<td>0.570</td>
<td>0.580</td>
<td>0.580</td>
<td>0.580</td>
</tr>
</tbody>
</table>

Average mean agreement between annotators, and between annotators and metrics.

Correlations between human judgments and metrics. Bold: highest correlation per language. * means significantly different from LAS.

### Analysis

- Preference of attachment or labeling for items where human and system disagree and human agreement ≥ .75.

### Examples

- UD
- CDT
- PDT