RRGparbank: A Parallel Role and Reference Grammar Treebank

Tatiana Bladier, Kilian Evang, Valeria Generalova, Zahra Ghane, Laura Kallmeyer, Robin Møllermann, Natalia Maors, Rainer Osswald, Simon Pettitjean

TreeGraSP (ERC) / Heinrich Heine University Düsseldorf (Germany)

first.last@hhu.de

RRGparbank in a nutshell

- Parallel corpus of RRG (Role and Reference Grammar) annotated sentences
- Source: George Orwell’s 1984 (≈ 6700 sentences) and its translations
- Languages: English, Farsi, French, German, Russian (+ Hungarian, Turkish)
- Largest RRG-based treebank available
- Genuine annotation, i.e. all sentences are manually validated
- Three classes of annotation:
  - Gold: validated by at least two annotators
  - Silver: validated by one annotator
  - Bronze: automatically generated parses
- 11 annotators from 2019
- Overall average annotator agreement of 95.3% EVALB f-score
- Multilingual data to browse, query, and download

Motivation

- Corpus-based investigations of RRG structures across languages thanks to query and download features
- Data-driven syntactic parsing for several languages
- Exploration of linguistic phenomena not yet covered by the RRG theory

First Release Statistics

- English, German, Farsi, French, and Russian seed data
- High degree of parallelism and a broad coverage of linguistic constructions

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>EN-seed</th>
<th>DE-seed</th>
<th>FR-seed</th>
<th>RU-seed</th>
<th>FA-seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tokens</td>
<td>122,843</td>
<td>23,750</td>
<td>23,444</td>
<td>24,670</td>
<td>17,097</td>
<td>22,456</td>
</tr>
<tr>
<td>Average sentence length</td>
<td>18.2</td>
<td>16.4</td>
<td>16.1</td>
<td>15.9</td>
<td>12.5</td>
<td>15.2</td>
</tr>
<tr>
<td>Not yet annotated</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1010</td>
</tr>
<tr>
<td>Gold with 1 annotation</td>
<td>2691</td>
<td>575</td>
<td>286</td>
<td>112</td>
<td>183</td>
<td>0</td>
</tr>
<tr>
<td>Gold with ≥ 2 annotations</td>
<td>3698</td>
<td>875</td>
<td>279</td>
<td>134</td>
<td>214</td>
<td>0</td>
</tr>
</tbody>
</table>

Selected Linguistic Phenomena

Discontinuous NUC for German particle verbs

<table>
<thead>
<tr>
<th></th>
<th>NP</th>
<th>NUC</th>
<th>CORE</th>
<th>PRO</th>
<th>PRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nothing</td>
<td>PRO</td>
<td>you</td>
<td>did</td>
<td>NP</td>
<td>not</td>
</tr>
<tr>
<td>happened</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Extraposed relative clause

<table>
<thead>
<tr>
<th></th>
<th>NUC</th>
<th>CORE</th>
<th>PRO</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Applications

Formalization of RRG: Tree Wrapping Grammar (TWG) [4,5,6]

<table>
<thead>
<tr>
<th></th>
<th>NUC</th>
<th>CORE</th>
<th>NUC</th>
<th>CORE</th>
<th>NUC</th>
<th>CORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>what</td>
<td>you</td>
<td>expect</td>
<td>me</td>
<td>to</td>
<td>say</td>
<td></td>
</tr>
</tbody>
</table>

Statistical parsing for several languages [7]

- Statistical TWG parser based on a fine-tuned multilingual BERT model and single-language BERT models
- Parser available at https://rrgparser.phil.hhu.de

Future Work

- Semantic annotations and frame-semantic parsing
- Finish annotation of non-English data

References