Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing

Željko Agić*, Daša Berović++, Danijela Merkler++, Marko Tadić+
*Linguistics Department, University of Potsdam, Germany, +Department of Linguistics, Faculty of Humanities and Social Sciences, ++University of Zagreb, Croatia
zagic@uni-potsdam.de, {dberovic, dmerkler, mtadic}@ffzg.hr

The research within the project XLike leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement nº 288342.

What?
• new version of the Croatian Dependency Treebank 2.0 (HOBS 2.0)
  – derived from the HOBS 1.0 (Tadić, 2007)
    • closely following PDT analytical layer annotation guidelines (Hajič et al., 2001)
    • number of issues with the manual annotation of subordinate clauses encountered
• departure from the original PDT annotation guidelines, but adapted better for Croatian

Where?
• subordinate clause annotation
• predicate in the subordinate clause is annotated with the syntactic tag of a syntactic element whose place it takes in the main clause
  – e.g., if subordinate clause = subject clause, its predicate is assigned with the syntactic tag Sub_Sb

How?
• new annotation guideline proposal
  – conjunctions in the tree should depend on the predicate in the main clause
  – predicate in the subordinate clause should depend on the conjunction that introduces it
• new syntactic sub-tagset for subordinate clause annotation in HOBS 2.0

Example of an attribute clause

- For adverbal clauses additional subclassification introduced

- Treebank statistics

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Tagset</th>
<th>Tokens</th>
<th>Sentences</th>
<th>MSDs</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOBS 1.0</td>
<td>basic</td>
<td>117369</td>
<td>4626</td>
<td>914</td>
<td>79.98</td>
</tr>
<tr>
<td></td>
<td>full</td>
<td>117369</td>
<td>4626</td>
<td>914</td>
<td>81.75</td>
</tr>
<tr>
<td>HOBS 2.0</td>
<td>basic</td>
<td>117369</td>
<td>4626</td>
<td>911</td>
<td>86.87</td>
</tr>
<tr>
<td></td>
<td>full</td>
<td>117369</td>
<td>4626</td>
<td>911</td>
<td>82.85</td>
</tr>
</tbody>
</table>

Experiment with parsing
• used graph-based MSTParser generator system (McDonald et al., 2005)
• backward compatibility with previous parsing experiments on HOBS 1.0 (Agić 2012, Agić and Merkler 2013, Agić et al. 2013)
• tenfold cross-validation parsing experiment

- Learning curves (LAS), two versions of HOBS