The Denoised Web Treebank
Evaluating Dependency Parsing under Noisy Input Conditions
Joachim Daiber and Rob van der Goot
University of Amsterdam, University of Groningen

OVERVIEW
- Novel benchmark for dependency parsing of noisy Web data.
- Our contributions:
  - Treebank
  - Evaluation of noise-aware parsing
  - Experiments

MAIN FINDINGS
- Text normalization improves parse quality on noisy content.
- Normalization works better above the word level.
- Treebank and evaluation metric: http://jodaiber.de/DenoisedWebTreebank

DATA
- 500 English Tweets randomly selected from 24h time window (07/01/2012).
- Manual language identification to avoid bias towards well-formed sentences.

TREEBANKS FOR NOISY CONTENT

<table>
<thead>
<tr>
<th>Name</th>
<th># Trees</th>
<th>OOV</th>
<th>Style</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>EWT [1]</td>
<td>16.4k</td>
<td>28%</td>
<td>C+D</td>
<td>No</td>
</tr>
<tr>
<td>Foster [2]</td>
<td>1k</td>
<td>25%</td>
<td>C</td>
<td>No</td>
</tr>
<tr>
<td>Foreebank [3]</td>
<td>1k</td>
<td>29%</td>
<td>C No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tweebank [4]</td>
<td>929</td>
<td>48%</td>
<td>D No</td>
<td>No</td>
</tr>
<tr>
<td>This work</td>
<td>500</td>
<td>31%</td>
<td>D Yes</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES