The Tembusu Treebank: An English Learner Treebank
Luis Morgado da Costa ♦, Francis Bond ♦, Roger V P Winder ♦

Introduction
We present the Tembusu Treebank — an open learner treebank created from the NTU Corpus of Learner English, unique for incorporating mal-rules in the annotation of ungrammatical sentences. We describe its development and evaluate it by training a new parse-ranking model for the English Resource Grammar, designed to help improve parse selection and grammatical error detection/diagnose.

Keywords: treebank, learner corpus, error detection, error diagnosis, parsing;

ERG & Mal-rules
• English Resource Grammar (ERG) is an open-source, broad-coverage computational grammar for English
• Uses Head-Driven Phrase Structure Grammar (HPSG) and produces Minimal Recursion Semantics (MRS)
• Coverage over unseen text between 81.2% and 96.8% (across a variety of genres)
• Incorporates substantial work on mal-rules – more than 270 different types
• Mal-rules expand prescriptive grammars
→ to selectively accept and identify ungrammatical sentences
→ can be used to trigger corrective feedback or to automatically correct a sentence through semantic reconstruction
• The ERG lacked a mal-rule enhanced parse-ranking model (limiting its usage)

Example: Sentence (1) requires multiple mal-rules to be parsed: i) the single noun error should not be able to form a bare noun phrase (NP); ii) the NP these rule should not be able to form due to agreement constraints;

iii) and, finally, if we assume a singular subject, there are also agreement issues between the subject and the main verb of the sentence.

(1) * These rule correct error.

NTU Corpus of Learner English
• The NTU Corpus of Learner English (NTUCLE) is an open corpus of learner English
• Comprising assignments from first year undergraduate engineering students from a major university in Singapore (NTU)
• Partially hand-tagged by six English lecturers (180 documents, 9,571 sentences)
• It currently contains around 800 documents (≈25,000 sentences)

The Tembusu Treebank
• The Tembusu Treebank hand tagged ≈20% of the NTUCLE (4,900 sentences) by five trained students, majoring in Linguistics and Multilingual Studies
• Who had to search for an adequate parse from all the parses generated by the ERG
• ≈35% (1700 sentences) were tagged and adjudicated by two or more people
• Quality confirmed by high levels of agreement (labeled: 73.1%, unlabeled: 78%)
• 76.3% of the 4,900 annotated sentences found a suitable tree, 890 sentences contained at least one mal-rule
• The treebank contains 1,253 mal-rule instances, distributed over 133 types

• It was used to train a new mal-rule enhanced parse ranking model for the ERG

Evaluation Experiments
• We used a test set of 1,000 unannotated sentences to evaluate the new parse ranking model (≈30 assignments)
• We compared the new mal-rule enhanced model with the original model with two configurations of the ERG: a simple set-up with mal-rules enabled (edERG); and a set-up with a filtering step (2-step)
• Tested the systems in two types of experiments: error detection and error diagnosis
• Systems using the new model performed better in both tasks (Tables 1 and 2)
• Boosts in Precision (17% in error detection, and 18% – 22% in error diagnosis) are especially relevant in the field of education

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERG (orig.)</td>
<td>1.000</td>
<td>0.007</td>
</tr>
<tr>
<td>edERG (orig.)</td>
<td>0.457</td>
<td>0.954</td>
</tr>
<tr>
<td>edERG (new)</td>
<td>0.627</td>
<td>0.834</td>
</tr>
<tr>
<td>2-step (orig.)</td>
<td>0.892</td>
<td>0.219</td>
</tr>
<tr>
<td>2-step (new)</td>
<td>0.892</td>
<td>0.219</td>
</tr>
</tbody>
</table>

Table 1: Error Detection Results

Conclusion
• Mal-rules are a suitable technology for error detection, correction and diagnosis – where explainability of results is crucial
• Learner treebanks offer new perspectives and opportunities for these tasks
• Results show very promising trends and motivate further work on this treebank
• Future work includes expanding the treebank size, addressing other limiting factors and exploring hybrid approaches using, e.g., a PCFG model trained from the Tembusu Treebank on top of the ERG

Release & Contact
The Tembusu Treebank will release all data (tagged and untagged) under a Creative Commons Attribution 4.0 International license. This data is available on Github, (https://github.com/Lmorgadodacosta/the-tembusu-treebank).
Contact: lmorgadodacosta@gmail.com

Acknowledgments
This research project received support from NTU through a Research Scholarship and an EdEx Teaching and Learning Grant administered by TLPD and from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-IF-2020 CHILL – No.101028782.

Language Resources and Evaluation Conference 2022