UNIB20005: Language and Computation

Project 3: Grammar Development

A context-free grammar is a set of rewriting rules, or productions, that models our intuitions about grammatical sentences. Starting with the grammar in `http://langtech.github.com/lac/projects/english.fcfg`, you will extend and test the grammar, and discuss your findings.

You may do this project individually, or in a team of two. (Team submissions should have wider coverage and more test cases, as specified in parentheses below.)

A grammar *test suite* is a collection of sentences that can be used to check that a grammar only accepts grammatical sentences. It contains a mixture of grammatical and ungrammatical sentences. A problem exists in the grammar if it rejects a grammatical sentence, or if it accepts an ungrammatical sentence. Remember that a sentence does not have to be meaningful in order to be grammatical (i.e. syntactic well-formedness is not the same as semantic interpretability). Obtain the test code from `http://langtech.github.com/lac/projects/gde.py`.

a. Create a test suite for the above grammar, in a file `sentences.txt`, containing at least 15 (25) grammatical sentences, and at least 15 (25) ungrammatical sentences, one per line. Use all lowercase, avoid punctuation, and separate each word with whitespace. Use words that are covered in the grammar. Mark ungrammatical sentences with an asterisk at the start of the line. For example:

   ```
   Jody claimed several cars disappeared
   *Jody walked several cars disappeared
   ```

b. Extend the grammar to support another kind of verb subcategorization, such as *put* (which requires an object noun phrase and a locative prepositional phrase), or *give* (which requires two noun phrases, or a single noun phrase followed by a prepositional phrase headed by *to*). Extend the test suite with at least 10 more grammatical sentences, and 10 more ungrammatical sentences. (Teams should do this twice, for two different verb types.)

c. Now extend the grammar with several more productions, by adding support for one (two) more syntactic construction(s), such as questions, relative clauses, cleft sentences, adverbial clauses (see the *SIL Glossary of Linguistic Terms* for explanations and examples). Take special care with your use of features. Use the lectures, and the syntax handout as a further source of ideas about possible syntactic constructions. Extend the test suite with at least 15 (30) more grammatical sentences, and 15 (30) more ungrammatical sentences.

Discuss your work in a plain text file named `report.pdf` or `report.txt` (∼ 400(750) words). Identify the team members in each file submitted. Your work will be assessed for correctness, clarity and insight. The report must be original. Your submission is worth 10% of the marks for this subject.

Note. The project has made simplifying assumptions relative to current approaches in "grammar engineering", which check that the expected parse trees and semantic representations were produced.

Please email your files (grammar, test sentences, report) with subject line *L&C Project 3* to Steven Bird at `sbird@unimelb.edu.au`, copying any team members. All submissions will be acknowledged. If you do not receive an acknowledgement within three days of the submission deadline, please resend.

Submit your work by the end of week 12 (10pm on Friday 26 October).