LING571 - Hw2
Chomsky Normal Form

1. Overview

Code for converting a non-CNF (Chomsky Normal Form) grammar to a CNF grammar was implemented using Python and the NLTK library.

In terms of file structure, there are 2 source files:

- Parse.py Defines the solution to Hw1.
- ConvertToCNF.py Defines the solution to Hw2.

2. Challenges

Honestly, the biggest challenge for this assignment was **learning Python**. This weekend, I learned everything from how __main__ is used as an entry point to the application to how type suggestion and lamdas work. It was a big learning curve; and getting to know the data structures used by the NLTK implementation of grammars took me well beyond a "Hello World" introduction to Python.

Otherwise, on a technical level, I found **eliminating unit productions** quite difficult. I think that I didn't originally understand that the objective was to remove the unit production by making it redundant through duplication. However, after watching a few YouTube videos and implementing the solution I was able to sort it out in my head.

3. Learning Outcomes

Certainly, I know a lot more about Python than I did a week ago!

Also, it was interesting to see that, while the conversion to CNF can be done relatively easily, algorithmically; it is much more difficult to do so efficiently; especially during the unit production phase, where my implementation consumes O(n) time for **each** unit production; putting the execution time at O(n^2) in the worst case.

4. Parse Comparison

Similarities

Looking at the parses generated by the original grammar and those from the CNF grammar, it's interesting to note that the average number of parses is exactly the same: 16.208! That is the most salient similarity.

Furthermore, this implies that (in all probability) each sentence has the same number of parses. To extrapolate further, the implication is that, while the CNF and non-CNF grammars are different, they can be applied to reach the same syntactic interpretations.

Differences

Of course, the two grammars are necessarily different. Specifically, they are defined by a different set of productions rules and, in particular, involve a different set of non-terminals (due to the dummy non-terminals created during the CNF conversion process). Therefore, the parses expressed in the two parse files are represented by different production rules.

5. Closing Comments

The assignment was pretty rough. But roughly on par for CLMS.