HACETTEPE UNIVERSITY

COMPUTER SCIENCE AND ENGINEERING

CMP670 – STATİSTİCAL NATURAL LANGUAGE PROCESSİNG

ASSIGNMENT-2

Mustafa KAYA

N18147271

1. Context-Free Grammars

A context-free grammar (CFG) is a set of recursive rewriting rules (or *productions*) used to generate patterns of strings.

A CFG consists of the following components:

- a set of *terminal symbols*, which are the characters of the alphabet that appear in the strings generated by the grammar.
- a set of *nonterminal symbols*, which are placeholders for patterns of terminal symbols that can be generated by the nonterminal symbols.
- a set of *productions*, which are rules for replacing (or rewriting) nonterminal symbols (on the left side of the production) in a string with other nonterminal or terminal symbols (on the right side of the production).
- a *start symbol*, which is a special nonterminal symbol that appears in the initial string generated by the grammar.

To generate a string of terminal symbols from a CFG, we:

- Begin with a string consisting of the start symbol;
- Apply one of the productions with the start symbol on the left hand size, replacing the start symbol with the right hand side of the production;
- Repeat the process of selecting nonterminal symbols in the string, and replacing them with the right hand side of some corresponding production, until all nonterminals have been replaced by terminal symbols.

1.1 LANGUAGE GENERATION WITH CFG

We can keep sentence lengths below a certain number in the program. In this study, sentences' length is under 30. The program consists of 5 sentences and these sentences are written to the file named “random-sentence.txt”. Some Generated Sentences:

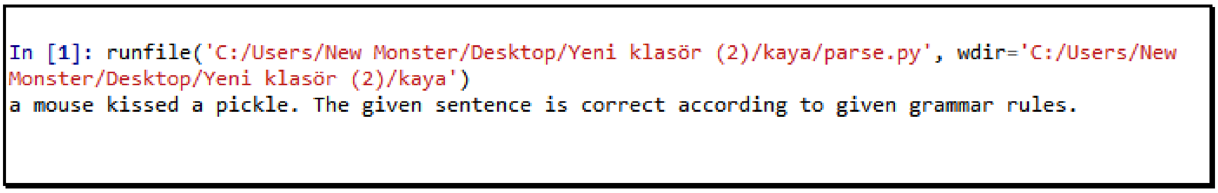
- ➢ every sandwich kissed a sandwich .
- ➢ a mouse kissed a pickle .
- ➢ every president with a floor on every delicious floor ate the president with every floor in every president !
- ➢ a floor ate the delicious floor on the pickle under every fine old mouse with a mouse on the mouse on a president on a pickle in a floor .
- ➢ the floor ate a sandwich !

It is observed that the types of sentences in which the NP structure is dominant are produced. Generated sentences are grammatically correct. But, given that, words are selected randomly, so most of them are meanless.

2. Parsing Sentences with CYK Parser

The Cocke–Younger–Kasami-Algorithm (CYK or CKY) is a highly efficient parsing algorithm for context-free grammars. This makes it ideal to decide the word-problem for context-free grammars, given in Chomsky normal form (CNF). Informally, the algorithm works as follows: In the first step write the word in the first row and add each non-terminal symbol in the row underneath which deduces the terminal symbols. After that, for each cell in the grid start vertically at the top and go down towards the cell to be checked and the second cell up in diagonal. For each such step, combine the cells and check if the combination appears in the grammar. If it does, add the left-hand-side non-terminal to the grid-cell. If after all steps the start-symbol is contained in the last row, the word can be derived by the given grammar.

The program code take sentence and checks whether the given sentence belong to the grammars and creates parsing trees. The result obtained by testing one sentence is given below. As it can be seen from the output, the program did not control the semantic structure, only the grammar control.



```
In [1]: runfile('C:/Users/New Monster/Desktop/Yeni klasör (2)/kaya/parse.py', wdir='C:/Users/New
Monster/Desktop/Yeni klasör (2)/kaya')
a mouse kissed a pickle. The given sentence is correct according to given grammar rules.
```

Figure 1: program output