**Course Project 2021/2022**

**Making Simple Lexical Analyzer and Parser for Natural Language Processing**

**Description:**

In this project, students (in group) need to define a Context Free Grammar (CFG) which represents rule in natural language. Based on that CFG, students make **a simple web based program (php/ java/ python/ .net, etc)**. The program acts as an lexical analyzer (to identify: does a lexical/token/word is valid according to defined terminal symbol) and parser (to check wether the order of the token/words comply with the rule of the Grammar).

This is an example of grammar which represents sentences in Bahasa Indonesia.

Grammar:

<S> ::= <noun> <verb> <noun>

<noun> ::= adik | kakak | buku | sepatu | bakso | tahu | topi

<verb> ::= membaca | makan | memakai

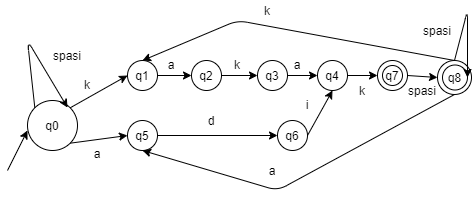
Non terminal symbol(s): S (starting symbol), noun, verb

Terminal symbol (s): adik, kakak, buku, sepatu, bakso, tahu, membaca, makan, memakai

Notes: BNF (Backus-Naur Form) notation is used to define the grammar

Based on the grammar above, this is Finite Automata to recognize lexical and the Parse Table:

1. Finite automata (Warning: This is incomplete since it only recognize two words. Students should make a complete FA for handling all terminal symbols)



1. Parse table LL(1)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **kakak** | **adik** | **bakso** | **tahu** | **membaca** | **makan** | **memakai** | **buku** | **sepatu** | **topi** | **EOS** |
| **S** | NN VB NN | NN VB NN | NN VB NN | NN VB NN | error | error | error | NN VB NN | NN VB NN | NN VB NN | error |
| **NN** | kakak | adik | bakso | tahu | error | error | error | buku | sepatu | topi | error |
| **VB** | error | error | error | error | membaca | makan | memakai | error | error | error | error |

Based on tge grammar and parse table, make program as the lexical analyzer and parser. Some test sentences will be used to check the program.

**Task (2-3 students/group):**

1. Make a simple CFG which represents rules/syntaxes in a particular human language ASIDE from Bahasa Indonesia, e.g.: Javanese, Sundanese, Balinese, Italian, Spanish, e.t.c. **It is recommended to choose language with the structure S-V-O (subject-verb-object), as listed on:** [**https://en.wikipedia.org/wiki/Subject–verb–object**](https://en.wikipedia.org/wiki/Subject–verb–object%23:~:text=Languages%20regarded%20as%20SVO%20include,(Indonesian%2C%20Malaysian)%2C%20Modern) because it has sama characteristic with our example. **Each group in class should choose a language that has not been choosen by the other group.** The number of non terminal symbol should be at least 3: starting symbol, and POSTag (word class) VB dan NN. The minimum number of terminal symbol (word) is 10.
2. Make FA based on CFG
3. Buatlah rparse-table LL (1)based onthe CFG.
4. Make a simple web based program.
5. Conduct tests to the program:
6. For lexycal analyzer, use 3 valid and 3 invalid words.
7. Test sentences contain 2-5 words. Use 3 correct sentences and 3 incorrect sentences.

**Deliverables:**

1. Report containing the desription of point 1,2,4,5 of the task [Maximum POINT: 60]
2. CFG Description and FA design should be submmited by the latest **Wednesday, June 8th 2022.**
3. Complete report should be submmitted by the latest Wednesday, **June 15th 2022.**
4. Program which is ready to run and its manual (explanation how to operate it) [Maximum POINT: 40]
5. Lexycal Analyzer program and its manual should be submmited by the latest **Wednesday, June 8th 2022.**
6. Parser program and its manual should be submmited by the latest **Wednesday, June 8th 2022.**

**Scoring bonus:**

1. Program is able to read user’s input (not using hardcode): 10 points
2. Integration of lexical analyzer and parser program (For each sentence test, the program only need to be be called once. The sentence is processed by lexycal analyzer and if it is valid, it will be directly processed by parser): 10 points.

**Example of program using Python (Desktop) :** <https://www.youtube.com/playlist?list=PLsQUw5vHd9DQHSL4WdsE2HK6iK1nI5ocM>