COMPARATIVE ANALYSIS OF LCP (LEFT-CORNER-PARSING) AND CYK ALGORITHM FOR CHECKING INDONESIAN STANDARD SENTENCE PATTERN

Sri Susanti[1]

[1] Teknik Informatika - Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail : srisusanti237@yahoo.com[1]

ABSTRACT

An author must pay attention to the pattern of the sentences in the language of the inscription, because the pattern of sentence will be influential in the process of delivery and recipient of the message. Based on the patterns of nature, get to know some indonesian sentence patterns (1) S-P, (2) S-P-O, (3) S-P-Pel, (4) S-P-K, (5) S-P-O-Pel, (6) S-P-O-Pel-K, (7) S-P-O-K, and (8) S-P-Pel-K. This is the underlying process of writing needs to pay attention to the pattern of sentence, so the need to research about the examination pattern of sentence.

The algorithms used in checking a sentence very many kinds, two of which are algorithms LCP (Left-Corner-Parsing) and algorithms CYK (Cocke-Younger-Kasami). The algorithm is the algorithm for parsing left corner combination of Top down Parsing and Bottom up Parsing. The process of the algorithm's Left Corner in the Bottom Up Parsing starts and ends in the Top Down. CYK algorithm using two dimensional array structure in examination. Any checked group said, then carried out his sentence structure examination. From the difference of the workings of the algorithm, it will be studied and compared the ability of each algorithm in examining raw language sentence patterns of Indonesia. Both of these algorithms is done so that the analysis of known accuracy from each of the algorithms in the check out raw phrase patterns.

The test results obtained by examination of raw sentence patterns of degree of accuracy of the CYK algorithm (Cocke-Younger-Kasami) by 65% with an error rate of 0.35 and algorithms left corner parsing by 60% with an error rate of 0.40. But the level of accuracy obtained is not yet big enough because the dictionary POST Tags used there is still a shortage that is still a lot of words that could not be identified.
Keywords: algorithm, Left Corner Parsing, Cocke-Younger-Kasami, Indonesian Language, Sentenece Patterns.

1. INTRODUCTION

A grammar must meet the criteria of scientific or empirical. Empirical is grammar must be scientifically proven by every person, in every place and at every time. The teaching function of the sentence is the standard knowledge taught in language classes [1]. The sentence that came from some basic sentence patterns are developed according to the needs. The sentence pattern of development should be based on formal grammar rules that apply, so there is a good reciprocal relationship and clearly among words or groups of words that make up the sentence.

Earlier research has ever done to check grammar parsing algorithm by using. Research ever done that is titled "Implemetasi Left Corner on designing Examination Application parses the Grammar in the Indonesian Language Sentence" by Vita Meriati Pandiangan. Research done just for handling errors in Indonesia language book discourse for elementary school children (elementary school). Checking is done on the basis of morphological errors, and the sentence that will be examined are already specified in the form of a sentence which is SPOK CFG (context free grammars). Results from research that has been done saying that algorithms Left Corner in Parsing the grammar checks can be run properly, that can generate the appropriate sentence pattern inspection with the pattern that has been created and saved in the database [2]. But in such research is not described in the specific measure used to assess the extent to which the implementation of algorithms left corner parsing grammar checks succeed in Indonesia of standard.

Other studies regarding the examination of the grammatical ever done that is by using the CYK algorithm (Cocke-Younger-Kasami) entitled "application of dynamic Program Algorithm in the Cocke-Younger-Kasami" by Inas by Luthfi. CYK algorithm on the study checked the validity of a strand is a symbol in the CFG with the use of the structure of a two-dimensional array. CYK algorithm is an implementation of dynamic program which is quite easy to implement despite the worst case complexity of O (n3). Results from research From the above studies say that each algorithm runs well in grammar checks. But in research-research that has been done there are no parameters or the size of the assessment

to look at the extent to which the implementation of these two algorithms are used in the examination of the grammar of Indonesia. Algorithm of Left Corner Parsing in both the examination of context-free grammar (CFG), while the Cocke-Younger-Kasami (CYK) can recognize sentences with good compliance patterns of sentences in the form CNF (Chomsky Normal Form). By doing a comparison analysis of the second algorithm will then be known workings and accuracy in precision inspection of raw sentence patterns of each algorithm. In order to further a better algorithm be used in the inspection of raw language sentence patterns of Indonesia.

Based on problems that have been presented, then in this study carried out a comparison between the algorithms of the Left Corner and CYK algorithms for Parsing (Cocke-Younger-Kasami) in recognizing the raw language of Indonesia sentences patterns to analyze whether the algorithm of the Left Corner better Parsing of the CYK algorithm (Cocke-Younger-Kasami) in terms of the level of accuracy in check pattern sentences standard.

1.1 Syntactic Indonesian Language

The syntax is the branch of linguistics that discussed the structuring and arrangement of words into larger units, called syntactic units, i.e. words, phrases, clauses, sentences, and discourse.

1.1.1 Function Syntactic

Syntactic function is a kind of boxes in the syntactic structure into populated category-specific category. Boxes were named subject (S), the predicate object (P), (O), complement (Kom), and description (Ket). However, in the practice of speaking in the sequence is not the same. A formal function of S and P should always adadalam each clause because both are inter-related in that it can be said, that S is part of the clause that marks what is stated by talks about S.

The object is part of the verb be predicate in the clause. Its presence is largely determined by the verb ketransitifan. This means that if verbs are transitive then the object will appear, but if the verbanya are not transitive so objects would not exist. There are two kinds of object that is the object of affective and effective. Affective objects are objects that are not a result of the deeds of the predicate. Otherwise effective object is the object that is the result of deeds predicate.

Complementary (comp) or complement is part of P P which makes verbs that become complementary. Its position is similar to the O, O perbedanya only if its existence is determined by the nature of the transitive verbanya. While the comp is not its existence is determined by a factor of ketransitifan, but by necessity to complement P. Elements of S, P, O and complement the core clause, whereas the element description is the outer core of the clause. It is because the position of the caption in the clause more flexible, meaning that it can be at the beginning or at the end of clause clause [6].

1.1.2 Syntactic Category

Syntactic category is a kind or type a word or phrase into the charger functions syntax. Syntactic category with regard to the term noun (N), verbs (V), ajektifa (A), adverbs (Adv), numeralia (Num), prepositions (Prep), conjunction (Konj), and pronominal (Pron). In this case N, V, and A is the main category, while the other is an additional category. The syntactic function of fillers can be either Word may also be phrases, so that in addition there is a noun phrase noun there anyway (FN), in addition to the word verb phrase verb there anyway (FV), and in addition there is also ajektifa word phrase ajektifa (FA). In addition on the side there is a category of words adverbialen there are also phrases adverbs (FAdv), next to the word category is also the phrase numeralia numeral (FNum), and next to the word category is also a preposition phrase preposisional (FProp).

A formal category N or FN fills the function of S and O on clauses or verbs. It can also fill a function P on the nominal clause. Category V or FV formally fill the functions P on clause verbs, and category A or FA fill function P on clause ajektifa [6]. Then, a formal syntactic functions of fillers can be seen in Table 1.1 below:

Table 1.1 Filler Syntactic function

S	P	O
N	N	N
FN	FN	FN
	V	
	FV	
	A	
	FA	

1.1.3 Phrase

The phrase is a unit of syntax that is composed of two or more words in a clause, which occupied the functions of syntax. Judging from the position of the second elemental, distinguished presence of koordinatif phrases and phrases subordinatif. Koordinatif the position of the phrase second elemental equal, whereas the phrase subordinatif the second position of the elemental not equal. Nothing serves as the core elements of a boss who called the phrase and there are based as a subordinate who is referred to as an additional explanatory phrase [6].
The phrases are divided into several groups:

1. Noun phrase koordinatif (FNK)
2. Noun phrase subordinatif (FNS)
3. Verb phrase koordinatif (FVK)
4. The phrase verb subordinatif (FVS)

5. The phrase koordinatif (FAK)
6. The phrasesubordinatif (FAS)
7. The phrase preposisional (Fprep)

1.1.4 Clause

A clause is a unit of syntax that is predictive. This means that inside the unit or the construction of a predicate, if there is no unit in the predicate, then it's not a unit clause. The position of the predicate is very important, because the types and categories of predicate that is what determines the presence of a function of the subject (S), the function of the object (O), complementary functions, and so on [6]. In functional analysis a clause are analyzed based on the functions of the elements into S, P, O, PEL, KET, and in the analysis of kategorial have been described that the charger function S consisting of N, the function P is composed of N, V, FP, Num, function O consists of N, the function is composed of MOP N, V, Num, and KET function consists of FP, N.

Based on categories that fill the P function can be differentiated into:

1. Clause verb
2. Noun Clause
3. Clause ajektifa
4. Clause preposisional
5. Clause numerial

1.1.5 Sentence

The sentence is a syntactic Unit assembled from its constituent base, which usually come with a conjunction of clauses, when needed, as well as accompanied by final intonation. Intonation is important in terms of final formation of a sentence can be either declarative intonation (ang in a language variety marked the point of writing), the interrogative intonation ("variety wrote given the question mark), the intonation of the imperative (" variety wrote was given an exclamation mark), and interjektif intonation ("variety wrote was given an exclamation mark). Without this final intonation of a clause would not be a sentence.

1.2 Chomsky Normal Form (CNF)

CNF is one normal form is very useful for the CFG has undergone simplification. Production rules in the form of his right right segment CNF form a terminal or two variables. In CNF, right section can only be either a terminal or two symbols the symbol variable. If there is more than one terminal symbol then to do the replacement and also if there are more than two symbol variable then to do the changes.

Example of production rules can be seen as follows:

A → BC
A → b
B → a
C → BA | d

1.3 POS Tag

The POST (Part-of-Speech) Tag is a designation of class words, such as nouns, verbs, adjectives, etc. POS Tagger is an application that is able to process the annotations part-of-Speech tags for each word in your document automatically. POS Tags are used as an aid in identifying token/tag each word taken from the post Tag Indonesia made by Arawinda Dinakaramani, Fam Rashel, Andry, Ruli Manurung and Admin [8].

POST Tags will identify which words including nouns, verbs, adjectives, adverbs, prepositions, conjunctions, pronouns, words and numbers. POST this Tag uses a Rule-Based approach based on the grammar rules of Indonesia. First of all, the post Tag will do the tokenisasi of the text using English-Indonesia. The next words that belong to the type of closed-class word processing. Then every ambiguous word processed using the rules defined for the class to find the right words.

1.4 Algorithm Left Corner Parsing (LCP)

Algorithm of left corner parsing is a combination of top-down parsing and bottom-up parsing is a strategy that is using data in bottom-up parsing and predictions from top-down parsing. The workings of this algorithm is to first accept a word constituent type, specify what starts with kind words. Will then be performed against the rest of the parsing process of constituent in top-down. Thus the process of parsing algorithm on left corner parsing starts in bottom-up and top-down are terminated [9].

In the process of vetting algorithm left corner parsing operations used are reduce, move, and remove. Used to reduce operating segment left the production rules from a terminal or variable. To move operations are used when the leftmost segment of categories there is the symbol $, where the symbol is the delimiter. The move operations namely moving from stack to stack the constituent categories. While the remove operation operation used to remove if there are variables or terminal on the same sentence and categories stack (stack on pop).

For example, there is a production rule as follows:

S → ASB | d
A → a
B → b

The steps of how to work the left corner parsing algorithms can be seen in table 3.2 as follows:

Table 1.2 The Step Left Corner Parsing

Sentence	Categories	Constituent	Operation
adb	S	ε	

db	$S	A	reduce A →a
Adb	S	ε	move
db	SB$S	S	reduce S →ASB
b	SBS	SS	reduce S →d
Sb	SB$S	S	move
b	B$S	S	remove
ε	BS	BS	reduce B →b
B	B$S	S	move
ε	S	S	remove
S	S	ε	move
ε	ε	ε	remove

1.5 Algorithm Cocke-Younger-Kasami (CYK)

CYK algotihm is an algorithm for parsing and or context-free grammar. CYK algorithm invented by j. Cocke, DH. Younger, and t. Kasami. Conditions for use of the CYK algorithm is grammar must be in the form of the CNF. The purpose of this algorithm is to indicate whether a string can be retrieved from a grammar [7].

The process of parsing the CYK algorithm with strand using data structures a two-dimensional array with the number of rows and columns of n x n. the value of n is determined from the number of words in a sentence. The process of parsing of the CYK algorithm utilizing the results of parsing to determine if the ongoing process is acceptable or not.

For example, there is a rule with variable initial production s.

S→AB | EC
F→CA
G→CD
H→EF
A→a
B→b
C→c
D→d
E→e

The string that will be checked for example "abcde". The first step is to create a two dimensional table as picture 1.1.

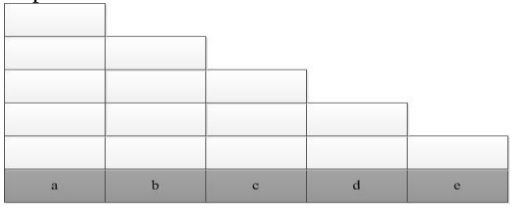


Figure 1.1 The establishment of the CYK Parsing Table

Each word in a string is placed on each block like image 1.1. Each block is filled based on the variables in accordance with the rules of production of CNF that was already established before. For example in the first row and the first column is filled with variables that can produce a, is the variable a. results can be seen in Figure 1.2.

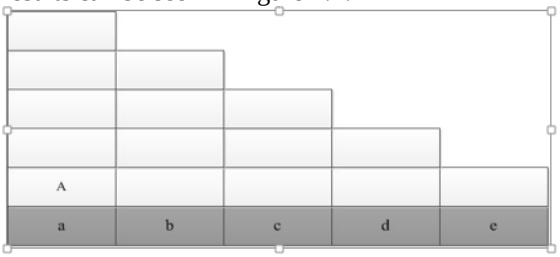


Figure 1.2 Filling the table with A variable

The next block is filled with the same steps, so that it can be seen the first line of its contents like images 1.3

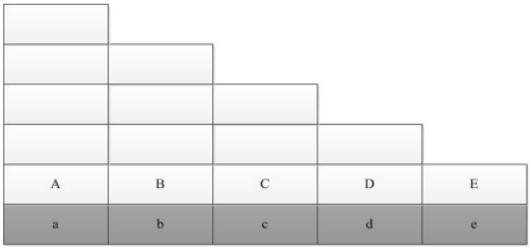


Figure 1.3 Filling The First Table

On the next line of how charging is done by looking at the variables in the two blocks, then views which produces both variable variables. As an example of the second line of the first column is filled with the sight of a variable on a first block and two blocks, i.e., variables A and B are included in the variable S, so how pengisiannya looks like in Figure 1.4.

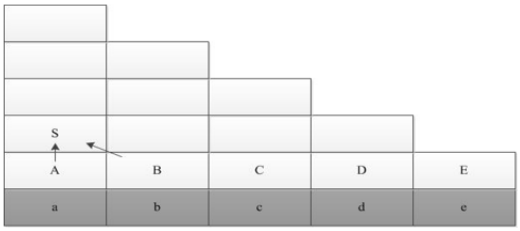


Figure 1.4 Filling the table with B variable

The next block is filled with the same steps in accordance with the blocks will be filled. When charging is done on the second block, the variable is taken from the second and third columns. For more details can be seen in Figure 1.5.

Figure 1.5 Charging Table Second Row

Charging on the third line and almost the same as the second line, but there is a difference from the way the pengecekkan is done. Filling the third row

from the first row to the third line before the (n-1). How pengecekkan can be seen as pictures 1.6.

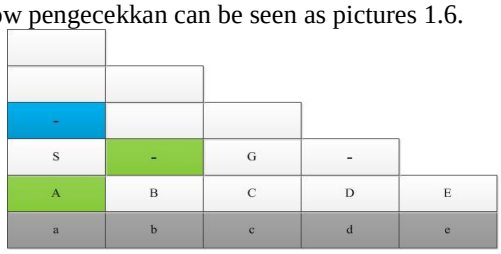


Figure 1.6 How To Check The First Step

The blue color indicates the blocks will be filled, while the green colour is checked from the blocks. In the second picture above the third row can be seen for how pengecekkan is done diagonally and vertically. How the next charging same as pictures 1.6. The end result of checking of the above expressions can be seen in Figure 1.7.

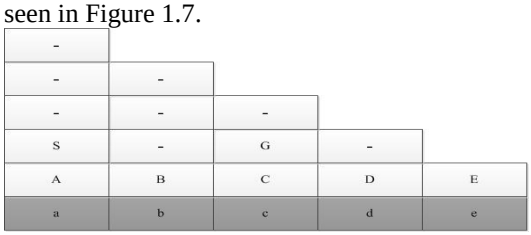


Figure 1.7 The results of Parsing the String with the CYK Algorithm

2. CONTENT OF RESEARCH

2.1 Analysis and Data Input

Data input to the simulator in the form of a news article with document formats. docx. Every sentence in the document are examined one by one and each word in the sentence is classified into the Group of words. Grouping words taken from the dictionary using the POS Tag Indonesia.

2.2 Syntactic Analysis for Production Rules

Based on a discussion of Indonesia language syntax and grouping words with POS Tag Indonesia before, then this section will explain how to make a production rule in the form of the CNF. Previously, the entire naming a group must be changed into the rules POS Tag so that it is not difficult in checking the production rules.
These changes can be seen in table 2.1.

Table 2.1 Naming a group into the POST Tags

Name Of The Groups	Change To
Noun / Kata Benda (N)	NN, NNP (FW), NND, PRP, atau WH
Adjective / Kata Sifat (A)	JJ
Verb / Kata Kerja (V)	VB
Adverb / Kata Keterangan (Adv)	RB
Numeralia (Num)	OD dan CD
Preposisi (P)	IN
Demonstrative (Dem)	PR
Konjungsi Koordinatif	CC
Artikula	DT
Verba Bantu	MD
Kata Ingkar	NEG
Partikel	RP
Konjungsi Subordinatif	SC
Simbol	SYM

2.3 Analysis Of Processing

Analysis of processing is the process flow describes the initial system will be built. The early stages of the system is the pengecekkan data input. The data input is allowed is formatted documents. docx. The system will then perform pre-launch stage processing, i.e. the document that has been entered will be broken down into the smallest unit (Word). Then the Word will be grouped into classes based on the words of the HEADING Tags, then it will be done the inspection process pattern sentences using both algorithms. After that will come out of results of examination of the algorithm, who calculated the level of its accuracy. Process design can be seen in Figure 2.1.

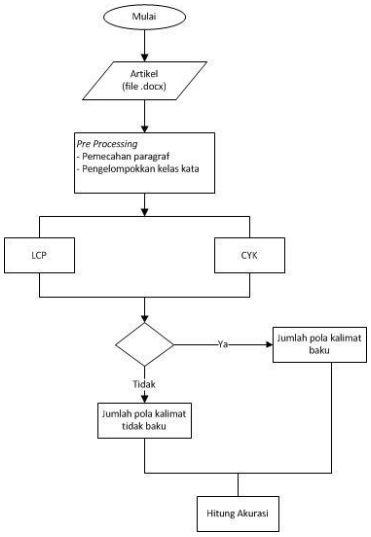


Figure 2.1 Flowchart System

2.4 Analysis Method

Application of analysis method of analysis is the method or algorithm in solving existing problems. In this research method is the algorithm the left corner and CYK algorithms for parsing (Cocke-Younger-Kasami).

2.4.1 Tahapan Algoritma *Left Corner Parsing*

There is a set of processes on algorithms left corner parsing is done in the examination of raw language sentence patterns of Indonesia, the process should be done on every stage, ordered produce a value as a result of the process, will then be processed and reused for the next phase, the algorithm processes the left corner parsing include:

1. Splitting paragraphs

In the early stages of document news articles will be checked pattern sentences baku will be broken down into paragrafnya sentence then the sentence being said. Splitting the paragraph became a sentence using regex. Regex is a technique used to match strings of text, such as characters, words, or patterns of characters. This stage will be known there are how many sentences in one paragraph of a news article that was tested. Point be a sign between sentences and be the final marker in the sentence. However, not all the point is a marker of the end of a sentence. Regex pattern formed is the [character] [the dot] [space]. The initial stage of checking is done by tracking the characters for the sake of existing characters at paragraph news articles online.

2. Word class grouping

At this stage each word grouped in accordance with the dictionary being taken based on rules of the heading tags. Table POS Tags can be seen in Appendix a. the comma at the end of the sentence should be included for the processes of formation of sentence patterns.

.

3. The review process raw sentence patterns

At this stage carried out an examination of whether the sentence patterns are examined in accordance with the rules of production or not.

A series of processes on the stage of the algorithm's left corner parsing can be seen in Figure 2.2.

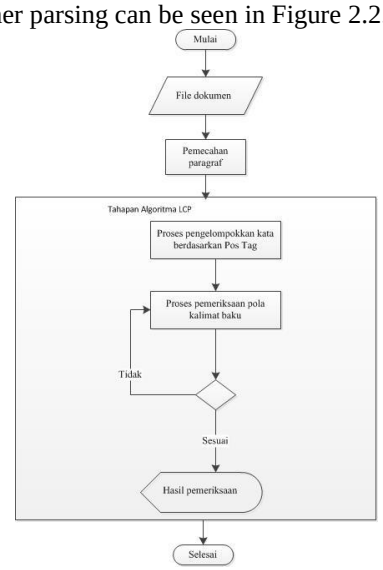


Figure 2.2 Flowchart of implementation Algorithms left corner parsing

2.4.2 Stages CYK Algorithm

CYK algorithm (Cocke-Younger-Kasami) there is also a series of processes that are carried out in the examination of raw language sentence patterns of Indonesia, the process should be done on every stage, ordered produce a value as a result of the process, will then be processed and reused for the next stage, the process of CYK algorithm (Cocke-Younger-Kasami) is almost the same as stages in the process of parsing algorithms left corner. CYK algorithm processes include:

1. Splitting paragraphs
2. Classification of words based on the POST Tags
3. Formation of sentence patterns into raw Tag

Processes on the stage of the CYK algorithm (Cocke-Younger-Kasami) can be seen in Figure 2.3.

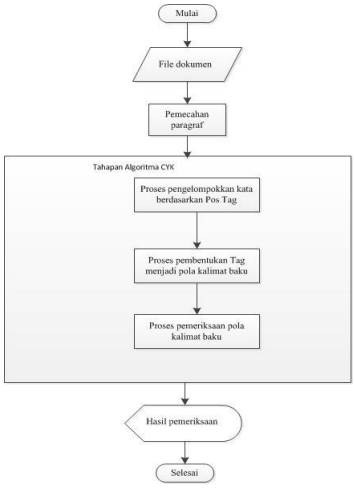


Figure 2.3 Flowchart Implementation Algorithm CYK

2.5 System Design

The design is a depiction, planning, and production sketches or settings of some separate elements into a unified whole and functioning.

2.5.1 Design Menu Structure

The design of the structure of the menu is the image path system usage so that the system that is built to be easy to understand and easy to use. The image design of the menu structure can be seen in Figure 2.4.

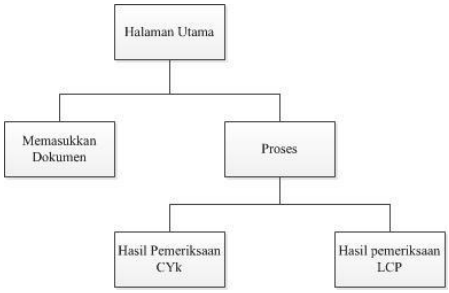


Figure 2.4 Design Menu Structure

2.5.2 Design Interface

The design of interface design is in determining the layout of the display system to be created. The design made as the design of interfaces on the system. The design of the interface is adapted to the users accessing the system, here is a design of the interface for form.

1. Main Page F01-Form Entry Program

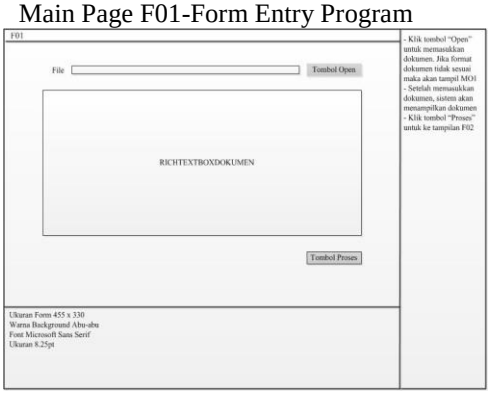


Figure 2.5 Form Main Page

2. F02-Form Examination Pattern Sentences

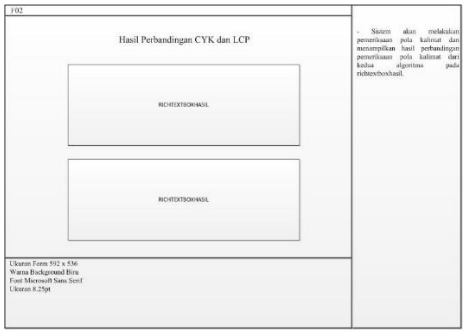


Figure 2.6 Form Examination Pattern Sentences

2.5.3 Implementation Of Interface

Interface implementations are implementations of the interface design has been created that is adapted to navigation along with the functionality of the page. Figure 2.7 and 2.8 is the Image display system interfaces for implementations of algorithms left corner CYK parsing algorithm for inspection of raw language sentence patterns of Indonesia that has been created.

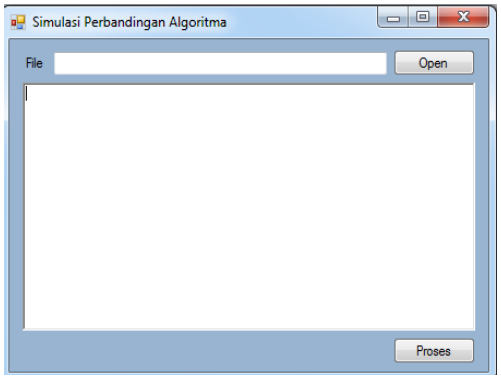


Figure 2.7 Interface Main Page

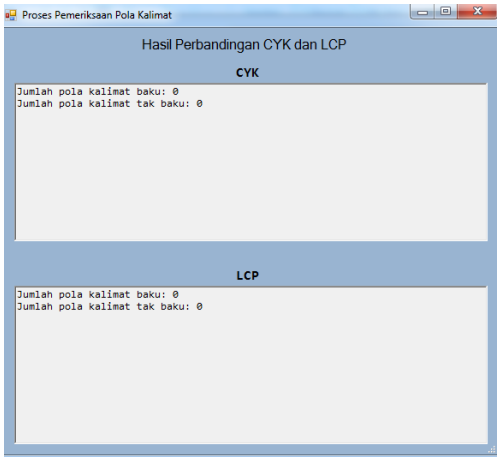


Figure 2.8 Interface Main The Result Examination Pattern Sectences

3 FINAL

Based on the results obtained in the research and drafting of this thesis as well as tailored to the purpose of the research, then the conclusion that algorithm CYK parsing algorithm left corner and can recognize a sentence properly in accordance with the rules of production which has been created. The more rules that created it will be more and more also a sentence pattern can be recognized. The test results of raw language sentence pattern inspection of Indonesia stated CYK algorithm has the level of accuracy of 65% with a rate of error of 0.35, while the left corner has a parsing algorithm accuracy rate of 60% with a rate of error of 0.40. The value obtained is not yet big enough because there is still a shortage of them POST the tags dictionary are used offline and still lacks many of the words are not yet known. To complete a portion of a word that does not exist in the dictionary author Tag POST using the great dictionary of the language of Indonesia (KBBI).

Based on all the processes in building this application advice that can be given in application development simulation examination pattern sentences baku, bahasa indonesia is as follows:

1 Applications this simulation can be developed again with the use of dictionaries online so that

Tags get POST data in realtime and the level of accuracy obtained will be greater.

2 Be able to use other methods or algorithms for comparison, in order to obtain a better algorithm is again in the process of examination of the raw language of Indonesia sentences patterns.

BIBLIOGRAPHY

[1] P. D. Iswara, "Penelitian Variasi Pola Kalimat," 24 februari 2009. [Online]. Available: http://iswara.staf.upi.edu/2009/02/24/penelitian-variasi-pola-kalimat. [Accessed 5 September 2015].

[2] . M. Pandiangan, "Implementasi Left Corner Parser Pada Perancangan Aplikasi Pemeriksaan Tata Bahasa Dalam Kalimat Bahasa Indonesia," vol. v, pp. 1-5, 2015..

[3] I. Luthfi, "Aplikasi Program Dinamis dalam Algoritma Cocke-Younger-Kasami (CYK)," pp. 1-5, 2007.

[4] N. S. Ibrahim, Penelitian dan Penilaian Pendidikan, Bandung: Sinar Baru, 1989.

[5] M. S. Rosa A. S, Rekayasa Perangkat Lunak, Bandung: BI-Obses, 2014.

[6] A. Chaer, Sintaksis Bahasa Indonesia (Pendekatan Proses), Jakarta: Rineka Cipta, 2009.

[7] F. Utdirartatmo, Teori Bahasa dan Otomata, Yogyakarta: Graha Ilmu, 2005.

[8] F. R. A. L. d. R. M. Arawinda Dinakaramani, "Pos Tag Indonesia," 20 Oktober 2014. [Online]. Available: http://bahasa.cs.ui.ac.id/postag/tagger. [Accessed 10 Oktober 2015].

[9] L. Kassner, "Left Corner Parsing," pp. 33-97, 2007.

[10] E. Prasetyo, "Klasifikasi," p. 6, 2012.