



UNIVERSITÄT  
BAYREUTH

University of Bayreuth  
Institute for Computer Science

## Bachelor Thesis

in Applied Computer Science

**Topic:** A Constrained CYK Instances Generator:  
Implementation and Evaluation

**Author:** Andreas Braun <[www.github.com/AndreasBraun5](https://www.github.com/AndreasBraun5)>  
Matrikel-Nr.: 1200197

**Version date:** August 30, 2017

**1. Supervisor:** Prof. Dr. Wim Martens  
**2. Supervisor:** Dr. Matthias Niewerth



## Abstract

Every year, lecturer in the field of theoretical computer science or an related one face the task to create an exam exercise that tests if their students have understood the way of working of the Cocke-Younger-Kasami algorithm. Various implementations and small online tools of the CYK algorithm can be found, but none actually assists during the process of creating an exercise.

Therefore various algorithms to generate specifically suitable exercises have been designed and compared through their success rates. The different approaches for these algorithms involve the uniformly randomly distribution of elements and the general Bottom-Up and Top-Down parsing approaches.

A GUI tool to automatically generate these exam exercises has been implemented. Its functionality contains that input parameters such as the count of variables, the count of terminals and the size of the word can be given. Suitable exam exercises are generated and one can be chosen for further modification and creation of the final exam exercise.

## Zusammenfassung

Jedes Jahr stehen Dozenten der theoretischen Informatik oder Dozenten eines verwandten Bereiches vor der Aufgabe Klausuraufgaben zu erstellen, um zu prüfen ob ihre Studenten die Arbeitsweise des Cocke-Younger-Kasami-Algorithmus verstanden haben. Verschiedene Implementierungen und kleinere Online-Tools des CYK-Algorithmus gibt es bereits, aber Keines unterstützt beim Prozess des Erstellens einer Aufgabe.

Verschiedene Algorithmen wurden zuerst entworfen, um genau passende Aufgaben zu generieren und wurden anschließend auch miteinander über ihre Erfolgsrate verglichen. Die unterschiedlichen Ansätze für die Algorithmen beinhalten das gleichmäßig zufällige Verteilen von Elementen und die allgemeinen Ansätze des Bottom-Up und Top-Down Parsings.

Es wurde ein GUI-Tool implementiert um automatisch Klausuraufgaben zu generieren. Die Funktionalität des Tools beinhaltet, dass Eingabewerte wie die Anzahl der Variablen, die Anzahl der Terminale und die Wortlänge gemacht werden können. Geeignete Klausuraufgaben werden automatisch generiert von denen Eine für weitere Modifikation und letztendlich für die Klausuraufgabenerstellung ausgewählt werden kann.

# Contents

<b>Abstract</b>	<b>4</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Motivation . . . . .	5
1.2 Context Free Grammar . . . . .	5
1.3 General approaches of parsing . . . . .	7
1.4 Data Structure Pyramid . . . . .	8
1.5 Cocke-Younger-Kasami Algorithm . . . . .	9
<b>2 Algorithms</b>	<b>11</b>
2.1 Sub Modules . . . . .	11
2.2 Dice rolling the distributions only . . . . .	13
2.3 Dice rolling and Bottom-Up variant one . . . . .	14
2.4 Dice rolling and Bottom-Up variant two . . . . .	17
2.5 Split Top-Down and fill Bottom-Up variant one . . . . .	20
2.6 Split Top-Down and fill Bottom-Up variant two . . . . .	26
2.7 Evaluation of Algorithms . . . . .	31
2.7.1 Success Rates . . . . .	31
2.7.2 Problem space exploration . . . . .	33
2.7.2.1 Comparison of the stopping criteria . . . . .	34
2.7.2.2 Picking the best algorithm . . . . .	35
<b>3 GUI Tool: CYK Instances Generator</b>	<b>36</b>
3.1 Overview GUI . . . . .	36
3.1.1 Working with the program . . . . .	36
3.2 Exam Exercise . . . . .	37
3.3 Scoring Model . . . . .	38
3.4 Parsing input with ANTLR . . . . .	39
3.5 Other Matters . . . . .	40
<b>Erklärung</b>	<b>42</b>

# 1 Introduction

## 1.1 Motivation

Every year, lecturer in the field of theoretical computer science or an related one face the task to create the 4-tuple exam *exercise* = (*grammar*, *word*, *parse table*, *derivation tree*) that tests if their students have understood the way of working of the Cocke-Younger-Kasami (CYK) algorithm. For it exercises need to be created which is a bit of a time consuming task.

Various implementations and small online tools of the CYK algorithm can be found <sup>1</sup> <sup>2</sup> <sup>3</sup> but none actually assists during the process of creating an exercise.

Therefore algorithms are needed to generate specifically suitable exercises with a high chance of success. Also a GUI tool, that allows automatic generation of the suitable exam exercises and further modification, is required and so a own solution is implemented.

## 1.2 Context Free Grammar

Firstly, we define a Context Free Grammar (CFG) as follows:

**Definition 1. Context Free Grammar (CFG)**

A CFG is a 4-tuple  $G = (V, \Sigma, S, P)$ :

- $V$  is a finite set of variables.
- $\Sigma$  is an alphabet
- $S$  is the start symbol and  $S \in V$ .
- $P$  is a finite set of rules:  $P \subseteq V \times (V \cup \Sigma)^*$ .

It holds:  $\Sigma \cap V = \emptyset$ .

Secondly, we define a CFG with restrictions (CFGR) as:

**Definition 2. CFG with restrictions (CFGR)**

A CFG  $G = (V, \Sigma, S, P)$  is a CFGR if:

- $P \subseteq V \times (V^2 \cup \Sigma)$ .

Throughout this thesis a grammar is always synonymous with Definition 2. Note that a CFGR is not necessarily in chomsky normal form (CNF) because it is still possible that there are unreachable variables – from the starting variable – or useless rules. For further convenience the following default values are always assumed in this thesis:

<sup>1</sup>CYK online tool: <http://lxmls.it.pt/2015/cky.html>

<sup>2</sup>CYK parser implementation: <http://jflap.org/tutorial/grammar/cyk/index.html>

<sup>3</sup>CYK algorithm implementation in Java: <https://github.com/ajh17/CYK-Java>

- $V = \{A, B, \dots\} =: Lhse.$
- $(V^2 \cup \Sigma) = \{AA, AB, BB, BA, BS, AC, \dots\} \cup \{a, b, \dots\} =: Rhse.$

A rule consists of a left hand side element ( $lhse \in Lhse$ ) and a right hand side element ( $rhse \in Rhse$ ).

**Example: lhse and rhse**

$lhse \rightarrow rhse$  applied to  $A \rightarrow c$  and  $B \rightarrow AC$  means that  $A$  and  $B$  are a  $lhse$  and  $c$  and  $AC$  are a  $rhse$ . Elements of  $V^2$  are often referred to as variable compounds.

For a word and a sub word Definition 3 holds:

**Definition 3. Word  $w$  and sub word**

- Word  $w$ :  $w = w_0 \cdot w_1 \cdot \dots \cdot w_j$  and  $w \in \Sigma^*$ .
- Sub word  $sw$  of a word  $w$ :  $sw = w_k \cdot \dots \cdot w_{l+k}$  where  $0 \leq i$  and  $l + k \leq j$ .

For a language and a language over a grammar Definition 4 holds:

**Definition 4. Language  $L$  and language  $L(G)$**

- Language  $L$ :  $L$  is a language over an alphabet  $\Sigma$  that is subset of  $\Sigma^*$ , that is a set of words over that alphabet.
- Language  $L(G)$ :  $L(G)$  is the language over a grammar  $G$ , that describes the set of words over that alphabet.

Moreover in the context of talking about sets, a set is always described beginning with an upper case letter, while one specific element of a set is described beginning with a lower case letter.

**Example: Upper Case Letter for sets**

Example: A "Pyramid" is a set consisting of multiple "Cell"s, whereas a *Cell* is again a subset of the set of variables "V". A "cellElement" is one specific element of a "Cell". (For further explanation behind this specific example see chapter 1.4)

### 1.3 General approaches of parsing

Next, the basic approach that may help finding a good algorithm is explained informally analogous to [1]. At first, parsing is described in general and afterwards its two characteristics are explained.

**Definition 5. Backward Problem = Parsing ( $w \in L(G)$ )**

Input:  $w$  and a grammar  $G$ .

Output:  $w \in L(G) \implies$  derivation  $d$ .

It is called parsing if there is a word  $w$  given and it is of interest to know if it is element of  $L(G)$ . Parsing is also the basis of the Cocke-Younger-Kasami algorithm.

After having defined what parsing in general is, it is important to know the two different ways of parsing, that will act as an idea provider for the algorithms.

**Bottom-Up parsing**

Bottom-Up parsing means to start parsing from the leaves up to the root node.

Actually, Bottom-Up parsing is the method used in the Cocke-Younger-Kasami algorithm, which fills the parse table from the "bottom up" [1].

Bottom-up parsing starts by recognizing the words smallest sub words before its mid-size sub words, and leaving the largest overall word as the last.

**Top-Down parsing**

Top-Down parsing means to start parsing from the root node down to the leaves.

"Top-Down parsing starts with the root node and successively applies rules from  $P$ , with the goal of finding a derivation of the test sentence  $w$ ." [1] (The so called test sentence is synonymous to an word  $w$ .)

## 1.4 Data Structure Pyramid

To be able to describe how the different algorithms work in a simpler way, the help data structure *Pyramid* is defined – note that *Pyramid* is a set and starts with upper case.

**Definition 6.** *Pyramid*

$Pyramid := \{Cell_{i,j} \mid i \in [0, i_{max}], j \in [0, j_{max,i}], i_{max} = |w|-1, j_{max,i} = i_{max}-i\}$   
 where  $Cell_{i,j} \subseteq \{(V, k) \mid k \in \mathbb{N}\}$  denotes the contents of the j'th cell in row i  
 and  $[i, j] := \{i, i+1, \dots, j-1, j\} \subseteq \mathbb{N}$ .

The cell  $Cell_{i_{max},0}$  is called the root of such a *Pyramid* and Figure 1.4 shows the visual representation of one.

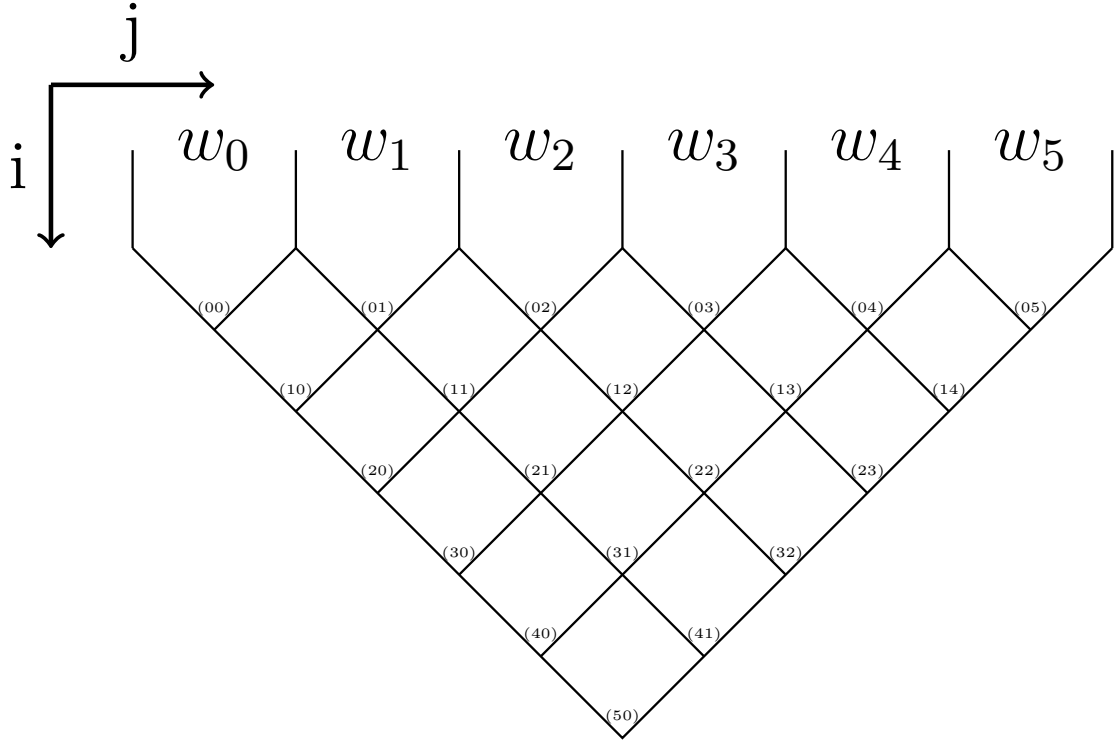


Figure 1: Visual representation of a *Pyramid* with the word  $w$  written above it.



## 1.5 Cocke-Younger-Kasami Algorithm

The Cocke-Younger-Kasami Algorithm (CYK) was independently developed in the 1960s by Itiroo Sakai [2], John Cocke and Jacob Schwartz [3], Tadao Kasami [4] and Daniel Younger [5].

The idea is to find all possible derivations of each sub word starting with size one and to consecutively use this information to find all possible derivations with a larger size of the subword up to the size of  $w$ . Finally it is checked whether  $w \in L(G)$  through the presence of the start variable in the root of the pyramid.

The description of the algorithm follows the source [6] adjusted to the data structure *Pyramid*. Later on it can be seen that the CYK algorithm can be used as a basis to find good algorithms.

Algorithm 1: CYK	
<b>Input:</b> Grammar $G = (V, \Sigma, S, P)$ and word $w \in \Sigma^* = \{w_0, w_1, \dots, w_j\}$ <b>Output:</b> $\text{true} \Leftrightarrow w \in L(G)$	
1	$Pyramid = \emptyset;$
2	<b>for</b> $j := 0 \rightarrow i_{max}$ <b>do</b>
3	$Pyramid = Pyramid \cup \{(X, j+1) \mid X \rightarrow w_j\};$ // Fills cells $Cell_{0,j}$
4	<b>end</b>
5	<b>for</b> $i := 1 \rightarrow i_{max}$ <b>do</b>
6	<b>for</b> $j := 0 \rightarrow j_{max,i}$ <b>do</b>
7	<b>for</b> $k := i-1 \rightarrow 0$ <b>do</b>
8	$Pyramid = Pyramid \cup \{(X, k+j+1) \mid X \rightarrow YZ, Y \in$ <span style="padding-left: 1.5em;"><math>Cell_{k,j}, Z \in Cell_{i-k-1, k+j+1}\};</math> // Fills cells <math>Cell_{i,j}</math></span>
9	<b>end</b>
10	<b>end</b>
11	<b>end</b>
12	<b>if</b> $(S, i) \in Cell_{i_{max},0}$ <b>then</b>
13	<b>return true;</b>
14	<b>end</b>
15	<b>return false;</b>
<hr/> Line 2: First row. Line 5: All rows except the first. Line 6: All cells in each row. Line 7: All possible cell combinations for each cell. Line 13: True if $Cell_{i_{max},0}$ contains the start variable.	

**Example: Algorithm CYK**

During the execution of the CYK algorithm the parsing table is filled as shown in Figure 2. At first the row with index  $i = 0$  is filled after Line 2 to Line 4 of the CYK algorithm, i.e. a  $Cell_{0,j}$  will contain the variable if it has the terminal  $w_j$  as its *rhse*. Then for each row  $i$  every cell with index  $j$  is looked at. Every possible combination of sub words for a cell are taken into account, i.e. for  $Cell_{4,1}$  there are the combinations of  $(Cell_{0,1}, Cell_{3,2})$ ,  $(Cell_{1,1}, Cell_{2,3})$ ,  $(Cell_{2,1}, Cell_{1,4})$  and  $(Cell_{3,1}, Cell_{0,5})$ . Applying Line 8 for example to the cell combination  $(Cell_{2,1}, Cell_{1,4})$  it leads to  $X \rightarrow AC$  here and because the compound variable  $AC$  is *rhse* of the variable  $S$  the  $Cell_{4,1}$  contains the element  $(S, 4)$ .

$A \rightarrow BS \mid a$   
 $B \rightarrow CB \mid b$   
 $C \rightarrow BC \mid c$   
 $S \rightarrow BA \mid AC$

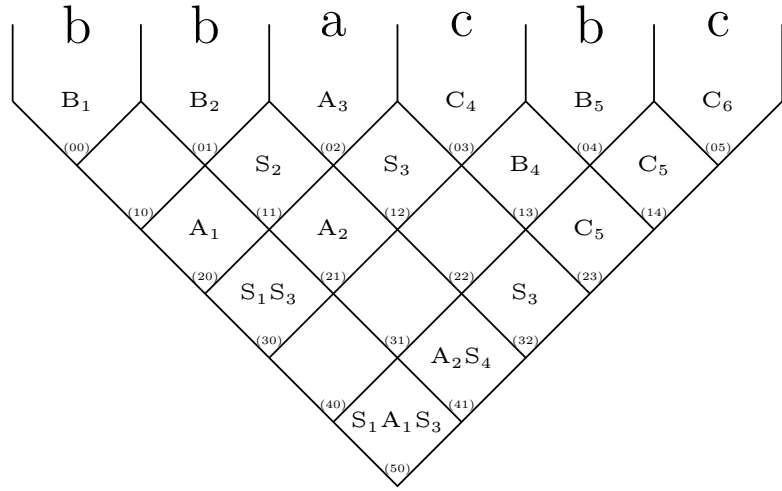


Figure 2: The CYK algorithm fills the cells of the pyramid during execution of Line 3 and Line 8.

## 2 Algorithms

### 2.1 Sub Modules

Sub modules are parts of the algorithms that are denoted circled with (A), (B), (C), (D) and (E). They are procedures that will be explained in more detail for a better understanding of the algorithms in the following chapters. (E) is explained not until Chapter 2.4 because it is needed only there.

**Distribute( $\Sigma, V$ ) (A) and Distribute( $V^2, V$ ) (B):**

The difference between (A) and (B) is that one time  $\Sigma$  and the other time  $V^2$  are distributed. But in both cases a uniform random subset of the *Rhse* is taken and again uniform randomly distributed over the set of available variables  $V$ . While distributing the terminals there exists at least one rule for every terminal used in the word  $w$ . The specifics of how they are distributed are described in the following algorithm:

<b>Algorithm 2:</b> Distribute	
<b>Input:</b> $V, Rhse \subseteq V^2$ or $Rhse \subseteq \Sigma$	
<b>Output:</b> Set of rules $P \subseteq V \times V^2$ or $P \subseteq V \times \Sigma$	
1	<b>foreach</b> $rhse \in Rhse$ <b>do</b>
2	<i>choose <math>n</math> uniformly randomly in <math>[i, j]</math>; // <math>i \in \mathbb{N}, j \in \mathbb{N}</math></i>
3	$V_{add} :=$ <i>uniform random subset of size <math>n</math> from <math>V</math>;</i>
4	$P = P \cup \{(v, rhse) \mid v \in V_{add}, rhse \in Rhse\};$
5	<b>end</b>
6	<b>return</b> $P$ ;

**Stopping Criteria (C):**

Two kinds of stopping criteria are used to determine whether an algorithm should terminate early on because an already suitable exercise has been found:

- stop if more than half of the pyramid cells are not empty any more
- stop if the root of the pyramid is not empty any more

Both stopping criteria are compared in Chapter 2.7 to see which one leads to more suitable exam exercises.

**CalculateSubsetForCell(Pyramid, i, j) (D):**

This procedure is needed to determine all possible compound variables out of all possible cell combinations for one specific cell. It works kind of analogous from Line 7 to Line 9 of the CYK algorithm (Algorithm 1).

<b>Algorithm 3:</b> CalculateSubsetForCell	
<b>Input:</b> $Pyramid, i \in \mathbb{N}, j \in \mathbb{N}$	
<b>Output:</b> $CellSet \subseteq V^2$	
1 $CellSet = \emptyset;$	
2 <b>for</b> $k := i - 1 \rightarrow 0$ <b>do</b>	
3     $CellSet = CellSet \cup \{YZ \mid X \rightarrow YZ, Y \in Cell_{k,j}, Z \in Cell_{i-k-1,k+j+1}\};$	
4 <b>end</b>	
5 <b>return</b> $CellSet;$	

**Example: Algorithm CalculateSubsetForCell**

In the following situation a rule is added to  $Cell_{3,0}$  while using Algorithm 3.

Grammar:

$A \rightarrow AB \mid a$

$B \rightarrow SC \mid b$

$C \rightarrow AB$

$S \rightarrow BA \mid AA$

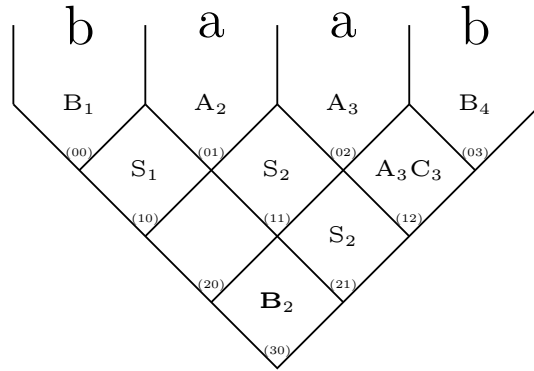


Figure 3: Example of Algorithm 3 while applying it on  $Cell_{3,0}$  via adding the rule  $B \rightarrow SC$ .

The calculation of  $CellSet$  for  $Cell_{3,0}$  results in  $\{SA, SC, BS\}$ , whereas  $SA$  and  $SC$  stem from  $Cell_{1,0}$  together with  $Cell_{1,2}$  and  $BA$  comes from  $Cell_{0,0}$  together with  $Cell_{2,1}$ . Now if either one of the rules  $lhse \rightarrow SA$ ,  $lhse \rightarrow SC$  or  $lhse \rightarrow BS$  is added to the grammar, then  $lhse \in Cell_{3,0}$ . Here the rule  $B \rightarrow SC$  has been added and finally  $(B, 2)$  is element of  $Cell_{3,0}$ .

In general if for one  $Cell_{i,j}$  a rule like  $lhse \rightarrow cs$  with  $cs \in CellSet$  (Line 3) is added, then automatically  $Cell_{i,j}$  will not be empty any more.

## 2.2 Dice rolling the distributions only

We start off by a primitive way of generating grammars, which will be the lower boundary while comparing the algorithms. Note that later on in Chapter 2.7.1 it is described what "performing better" means in the context of this thesis.

<b>Algorithm 4:</b> DiceRollOnlyCYK	
<b>Input:</b>	Word $w \in \Sigma^*$
<b>Output:</b>	Set of rules $P$
1	$P = \emptyset; \quad // \quad P \subseteq V \times (V^2 \cup \Sigma)$
2	$P = P \cup \text{Distribute}(\Sigma, V); \quad \textcircled{\text{A}}$
3	$P = P \cup \text{Distribute}(V^2, V); \quad \textcircled{\text{B}}$
4	<b>return</b> $P$ ;

The algorithm DiceRollOnly (Algorithm 4) distributes terminals  $\Sigma$  to at least one  $lhse$ , but a compound variable  $vc \in V^2$  does not has to be distributed at all. Note that for each terminal of  $\Sigma = \{a, b\}$  at least one rule like  $lhse \rightarrow a$  and  $lhse \rightarrow b$  is generated.

### Example: Algorithm DiceRollOnlyCYK

For each possible compound variable  $V^2 = \{AA, AB, AC, AS, BB, BC, BS, CC, CS, SS\}$  it is possible that only a smaller subset like  $\{AA, BA, CC, SC\}$  is distributed (Figure 4) so that only rules like  $lhse \rightarrow AA$ ,  $lhse \rightarrow BA$ ,  $lhse \rightarrow CC$  and  $lhse \rightarrow SC$  exist.

Grammar after Line 2:	Grammar after Line 3:
$C \rightarrow a$	$C \rightarrow BA \mid AA \mid a$
$B \rightarrow b$	$B \rightarrow b$
	$S \rightarrow CC \mid SC$

Figure 4: Shortend overview of the example of Algorithm 4.

### 2.3 Dice rolling and Bottom-Up variant one

Another approach to design an algorithm is after the Bottom-Up approach (Chapter 1.3) in which the parsing table is filled starting from the leaves in direction of the root node.

The basic idea is to guide the choice of rules while distributing the compound variables  $V^2$ . In Algorithm 4, the naive approach, it is possible that the terminals are distributed to the variables  $A$  and  $B$  and Algorithm 4 completely discards this fact during the distribution of the compound variables (see Figure 6 in the middle).

**Example: Disregarding already added rules**

Figure 5 shows the starting situation for this example.

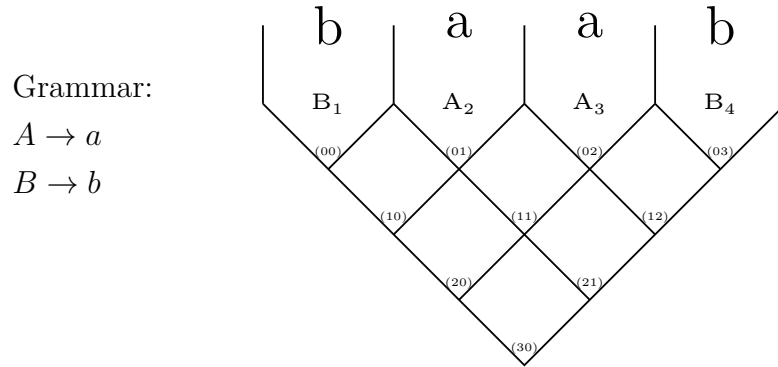


Figure 5: Disregarding the already added rules: Starting situation.

If rules like  $lhse \rightarrow CC$  or  $lhse \rightarrow SC$  are added they do not directly help to fill the parsing table and bloat the grammar with useless rules (see Figure 6). This is a unfortunate adding of rules that does not help to fill the parsing table and can happen in Algorithm 4.

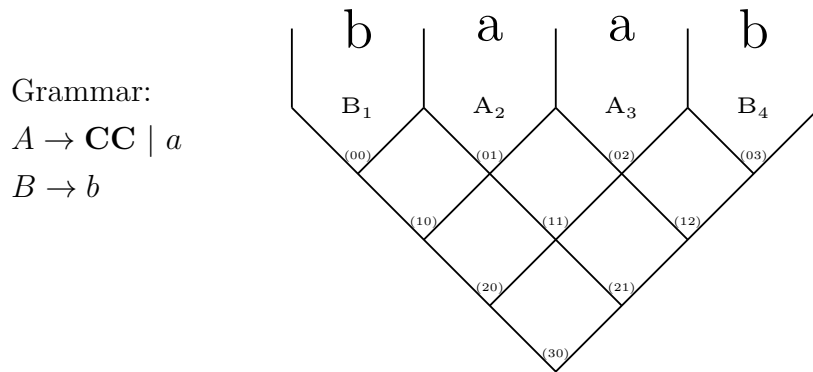


Figure 6: Disregarding the already added rules middle: Unfortunate adding. Bottom: Advantageous adding of rules.

More reasonable rules to add would be  $lhse \rightarrow BA$ ,  $lhse \rightarrow AA$  or  $lhse \rightarrow AB$  (see Figure 7). This is an advantageous adding of rules as intended in Algorithm 5 that helps to fill the *pyramid*.

Grammar:

$$A \rightarrow a$$

$$B \rightarrow \mathbf{AB} \mid b$$

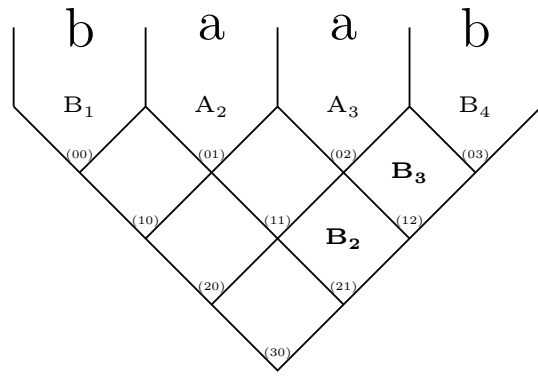


Figure 7: Disregarding the already added rules Bottom: Advantageous adding of rules.

Algorithm 5 continues on this idea: After distributing the terminals (Line 2) the updated parsing table (Line 12) is always taken into consideration while calculating (Line 10) variable compounds and to finally add a part of them (Line 11) in form of rules to the grammar. As explanation, for each chosen cell a *CellSet* (Line 10) is calculated, that only contains reasonable variable compounds. This way only variable compounds are added that directly help to fill the parsing table.

<b>Algorithm 5:</b> BottomUpDiceRollVar1	
<b>Input:</b> Word $w \in \Sigma^*$	
<b>Output:</b> Set of rules $P$	
1	$P = \emptyset; \quad // \quad P \subseteq V \times (V^2 \cup \Sigma)$
2	$P = \text{Distribute}(\Sigma, V); \quad \textcircled{\text{A}}$
3	$\text{Pyramid} = \text{CYK}(G, w);$
4	<b>for</b> $i := 1$ <b>to</b> $i_{\max}$ <b>do</b>
5	$J = \{0, \dots, j_{\max} - 1\}; \quad // \quad J \subseteq \mathbb{N}$
6	$\text{CellSet} = \emptyset; \quad // \quad \text{CellSet} \subseteq V^2$
7	<b>while</b> $ J  > 0$ <b>do</b>
8	choose one $j \in J$ uniform randomly;
9	$J = J \setminus \{j\};$
10	$\text{CellSet} = \text{CalculateSubsetForCell}(\text{Pyramid}, i, j); \quad \textcircled{\text{D}}$
11	$P = P \cup \text{Distribute}(\text{CellSet}, V); \quad \textcircled{\text{B}}$
12	$\text{Pyramid} = \text{CYK}(G, w);$
13	<b>if</b> stopping criteria met $\textcircled{\text{C}}$ <b>then</b>
14	<b>return</b> $P;$
15	<b>end</b>
16	<b>end</b>
17	<b>end</b>
18	<b>return</b> $P;$
Line 3: Fills the $i=0$ row of the pyramid. Line 9: A cell is visited only once.	



## 2.4 Dice rolling and Bottom-Up variant two

While examining Algorithm 5 via its log file (Figure 8) it can be seen that already a very small number of rules in the grammar is sufficient so that the stopping criteria  $\textcircled{C}$  is met – the cells that indirectly decide what rules to add are mostly from row one ( $i = 1$ ) and sometimes if at all from row two ( $i = 2$ ).

Final cell worked with Index: 1,2  
 Final cell worked with Index: 1,0  
 Final cell worked with Index: 1,6  
 Final cell worked with Index: 1,0  
 Final cell worked with Index: 1,2  
 Final cell worked with Index: 1,3  
 Final cell worked with Index: 2,4

Figure 8: Overview of log files of Algorithm 5 with  $|V| = 4$  and  $|\Sigma| = 2$ .

This again leads to a further improvement idea to introduce a row dependent *threshold<sub>i</sub>* (Line 9 of Algorithm 6 BottomUpDiceRollVars) which helps that more cells with  $i \geq 2$  are chosen – what possibly leads to more diverse grammars being generated. The diversity, in context of the procedure BottomUpDiceRollVar1 (Algorithm 5), is somewhat too restricted to the *lhse*s that have one of the terminals as its *rhse*. Most of the rules that are part of the grammar will contain one of these *lhse*s as explained in Chapter 2.3. This is caused by the basic idea of Algorithm 5 but also due to the relatively small number of rules that are added to the grammar altogether.

Further diversification is achieved through the usage of  $\textcircled{E}$  (Line 10 of Algorithm 6 BottomUpDiceRollVars), i.e. the variable compounds that already have been used in a row with low index  $i$  are at a disadvantage to be picked again (A more detailed explanation is found at the end of this chapter).

### Example: Better diversity in a grammar

As seen in Figure 9 the rules with  $BA$  and  $AA$  are added to the variables  $B$  and  $A$  in Grammar1. For Grammar2 instead the rule  $B \rightarrow SS$  is added that contributes to a better diversity compared to Grammar1. Grammar2 contains one more unique *rhse* ( $SS$ ) compared to Grammar1.

Grammar0:	Grammar1:	Grammar2:
$C \rightarrow BA \mid AA \mid a$	$C \rightarrow BA \mid AA \mid a$	$C \rightarrow BA \mid AA \mid a$
$B \rightarrow b$	$B \rightarrow BA \mid AA \mid b$	$B \rightarrow SS \mid b$
$S \rightarrow CC \mid SC$	$S \rightarrow BA \mid AA \mid CC \mid SC$	$S \rightarrow CC \mid SC$

Figure 9: Better diversity.: Starting point is Grammar0 and Grammar2 is of better diversity than Grammar1.

**Algorithm 6:** BottomUpDiceRollVar2

**Input:** Word  $w \in \Sigma^*$   
**Output:** Set of rules  $P$

```

1  $P = \emptyset$ ; //  $P \subseteq V \times (V^2 \cup \Sigma)$ 
2  $RowSet = \emptyset$ ; //  $RowSet \subseteq \{(xy, i) \mid x, y \in V \wedge i \in \mathbb{N}\}$ 
3  $P = Distribute(\Sigma, V)$ ; (A)
4  $Pyramid = CYK(G, w)$ ;
5 for  $i := 1$  to  $i_{max}$  do
6   for  $j := 0$  to  $j_{max} - i$  do
7      $RowSet = RowSet \cup \{(xy, i) \mid xy \in$ 
        $CalculateSubsetForCell(Pyramid, i, j)$  (D);
8   end
9   while  $threshold_i$  not reached do
10    choose one  $(xy, i)$  from  $RowSet$  uniform randomly with
      probability depending on  $i$ ; (E)
11     $P = P \cup Distribute(xy, V)$ ; (B)
12     $Pyramid = CYK(G, w)$ ;
13    if stopping criteria met (C) then
14      return  $P$ ;
15    end
16  end
17 end
18 return  $P$ ;

```

Line 4: Fills the  $i=0$  row of the pyramid.

**Choose one  $xy$  from  $(xy, i) \in RowSet$  uniform randomly with probability depending on row  $i$  (E) :**

At some point a decision needs to be made about what rule  $lhse \rightarrow xy$  with  $xy \in V^2$  will be added to the grammar. Depending on which  $xy$  is chosen the influence on the entire pyramid varies. Some  $xy$  only change the parsing table in one of its later rows ( $i \gg 1$ ) but other  $xy$  even change it in one of the first rows. If there is a change in one of the first rows it is more likely that the entire pyramid will be filled with more elements. Now the task of choosing rules to add, that only change the pyramid in one of the later rows, with a higher probability than the others is tackled with (E).

The approach here only makes sense together with (D) in which all possible compound variables are calculated that help to fill one specific cell.  $RowSet \subseteq \{(xy, i) \mid x, y \in V \wedge i \in \mathbb{N}\}$  whereas the  $xy$  are calculated with (D) and  $i$  is the row number of the specific cell.

**Example: Procedure (E)**

With *RowSet* the choice can be influenced regarding the row number  $i$ : Firstly the *RowSet* is compressed, i.e. every tuple with the same  $xy$  will be merged to its lowest  $i$ , as following:  $RowSet = \{(AB, 3), (AB, 1), (AB, 5), \dots\}$  will become  $RowSet = \{(AB, 1), \dots\}$ . Afterwards all elements of *RowSet* will be placed in the *RowMultiSet* that can contain multiple equivalent elements. Now each element of *RowMultiSet* will be weighted according to their  $i$ . That means that elements like  $(AB, 1)$  will only occur one time while elements like  $(BC, 3)$  will occur three times and so on:  $RowMultiSet = \{(AB, 1), (BC, 3), \dots\}$  becomes  $RowMultiSet = \{(AB, 1), (BC, 3), (BC, 3), (BC, 3), \dots\}$ . Now one element will be chosen uniformly randomly out of this weighted *RowMultiSet*. In the example in Figure 10 this results in  $xy = BC$ .

```

RowSet = {(AB, 3), (AB, 1), (AB, 5), ...}           // compress
RowSet = {(AB, 1), ...}                             // put in RowMultiSet
RowMultiSet = {(AB, 1), (BC, 3), ...}                // weight elements
RowMultiSet = {(AB, 1), (BC, 3), (BC, 3), (BC, 3), ...} // pick element
xy = BC

```

Figure 10: Overview of the procedure E.

## 2.5 Split Top-Down and fill Bottom-Up variant one

Until now we have only discussed algorithms that purely use the Bottom-Up approach, so another way is to utilize the Top-Down approach in combination with the Bottom-Up approach.

The idea here is first to distribute the terminals (Line 2 of Algorithm 7 SplitThenFill-CYK) and then to uniformly randomly generate a predefined structure of the derivation tree (Line 4 of Algorithm 7 and in general Algorithm 8 SplitThenFillCYKRec) Top-Downwards and then again to fill the parsing table Bottom-Upwards accordingly to fill this derivation tree. The structure of the derivation tree for instance can look as follows:

**Example: Derivation structure in the pyramid and as a derivation tree**  
The numbers correspond to the depth in the tree.

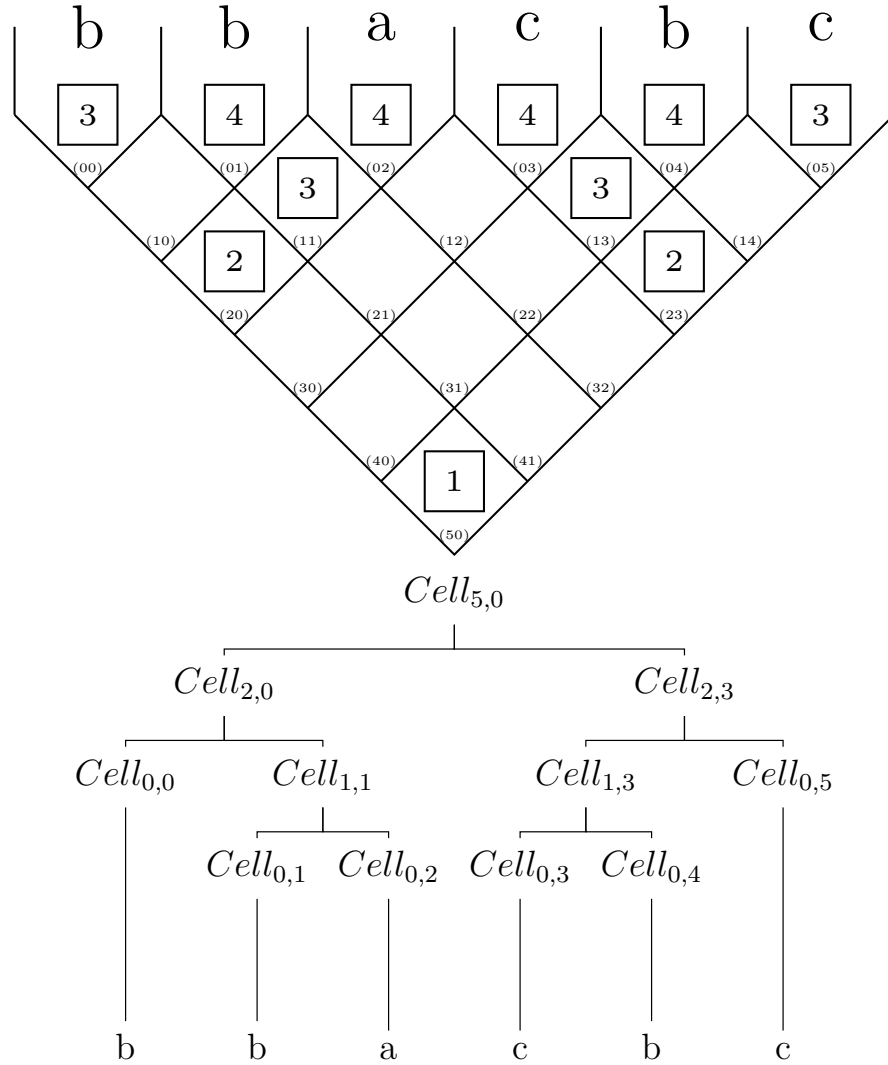


Figure 11: Derivation structure shown in the pyramid in the top and below as a derivation tree.

As the name of the algorithms implies only after completely generating the structure of the derivation tree (splitting of the word in sub words) the rules are added to the grammar that help filling the cells occurring in the derivation tree.

Now every time before adding a new rule (Algorithm 8 SplitThenFillCYKRec Line 15) the already available information regarding the other rules is used to determine if a new rule is needed to fill this node of the derivation tree (Line 12 of Algorithm 8 SplitThenFillCYKRec).

<b>Algorithm 7: SplitThenFillCYK</b>	
<b>Input:</b> Word $w \in \Sigma^*$	
<b>Output:</b> Set of rules $P$	
1	$P = \emptyset; \quad // \quad P \subseteq V \times (V^2 \cup \Sigma)$
2	$P = Distribute(\Sigma, V); \quad \textcircled{A}$
3	$Sol = (P_{Sol}, Cell_{i_{max},0}); \quad // \quad P_{Sol} \subseteq P \quad \wedge \quad Cell_{i_{max},0} \in Pyramid$
4	$Sol = SplitThenFillCYKRec(P, w, i_{max}, 0);$
5	<b>return</b> $P_{Sol};$
Line 2: Fills the $i=0$ row of the pyramid.	

For this algorithm it is important to mention that while using  $\textcircled{B}$  (Line 15 of Algorithm 8) a variable compound is added to at least one  $lhse$ . For every element of  $vc \in VarComp$  (Line 14 of Algorithm 8) there exists at least one rule  $lhse \rightarrow vc$ .

**Algorithm 8:** SplitThenFillCYKRec**Input:**  $P_{in} \subseteq V \times (V^2 \cup \Sigma)$ ,  $w \in \Sigma^*$ ,  $i, j \in \mathbb{N}$ **Output:**  $(P, Cell_{i,j})$ 

```

1  $P = P_{in};$ 
2 if  $i = 0$  then
3   | return  $(P, Cell_{i,j});$ 
4 end
5 choose one  $m$  uniform randomly in  $[j + 1, j + i];$ 
6  $(P, Cell_l) = SplitThenFillCYKRec(P, w, (m - j - 1), j);$ 
7  $(P, Cell_r) = SplitThenFillCYKRec(P, w, (j + i - m), m);$ 
8  $Pyramid = CYK(G, w);$ 
9 if stopping criteria met  $\textcircled{C}$  then
10  | return  $(P, Cell_{i,j});$ 
11 end
12 if  $Cell_{i,j} = \emptyset$  then
13   |  $VarComp = \text{uniform random subset of } \{vc \mid v \in Cell_l \wedge$ 
14     |  $c \in Cell_r\}$  with  $|VarComp| \geq 1;$ 
15   |  $P = P \cup Distribute(VarComp, V);$   $\textcircled{B}$ 
16 end
17 return  $(P, Cell_{i,j});$ 

```

---

Line 8: Needed to update the *pyramid* to be able to check if the stopping criteria is met in Line 9.



The next recursion step happens in  $Cell_{2,0}$ . Now  $Cell_l = \{B_1\}$  and  $Cell_r = \{S_2\}$ . Analogously the rule  $A \rightarrow BS$  is added to the grammar as seen in Figure 14:

Grammar:

$A \rightarrow \mathbf{BS} \mid a$

$B \rightarrow b$

$C \rightarrow c$

$S \rightarrow BA$

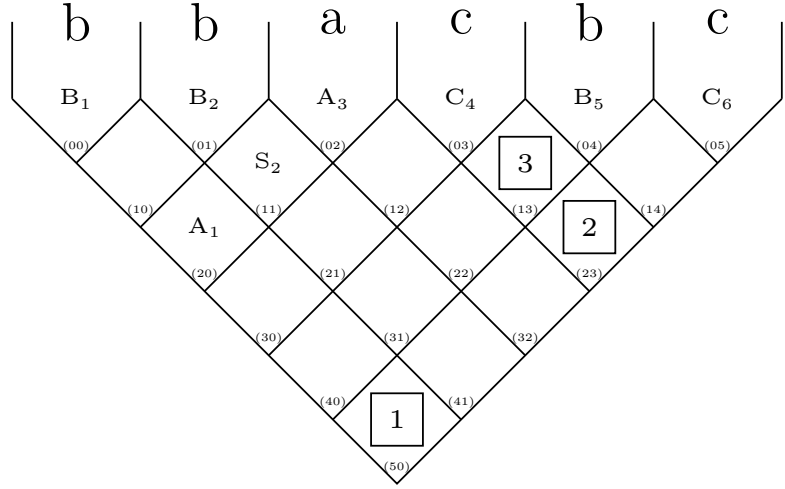


Figure 14: Illustration of Algorithm 7 SplitThenFillCYK part 3 after adding the rule  $A \rightarrow BS$ .

The next two analogous steps are described with Figure 15 and with Figure 16.

Grammar:

$A \rightarrow BS \mid a$

$B \rightarrow \mathbf{CB} \mid b$

$C \rightarrow c$

$S \rightarrow BA$

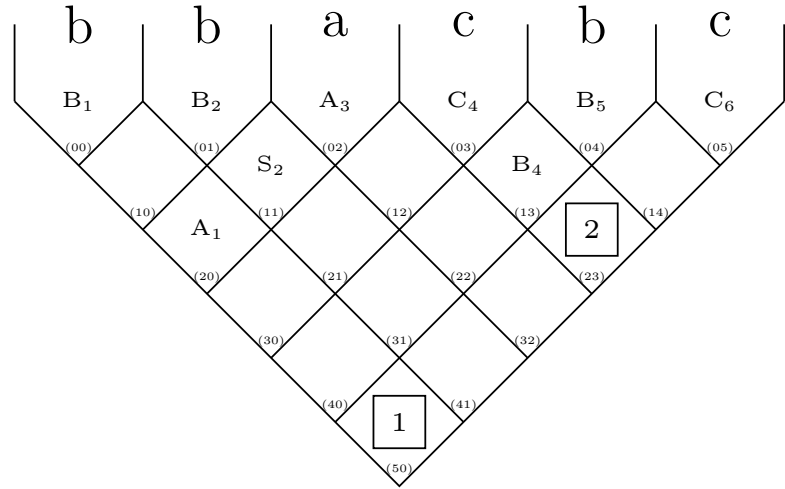


Figure 15: Illustration of Algorithm 7 SplitThenFillCYK part 4. The recursion step in  $Cell_{1,3}$  is resolved by adding the rule  $B \rightarrow CB$ .



Grammar:

$A \rightarrow BS \mid a$

$B \rightarrow CB \mid b$

$C \rightarrow \mathbf{BC} \mid c$

$S \rightarrow BA$

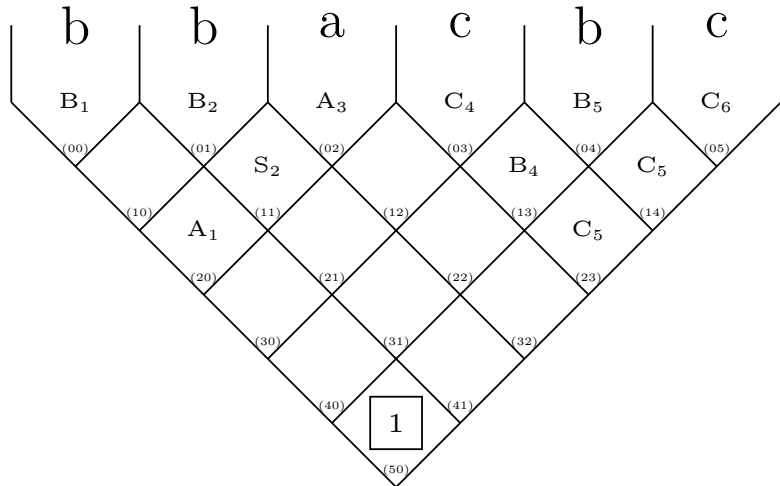


Figure 16: Illustration of Algorithm 7 SplitThenFillCYK part 5. The recursion step in  $Cell_{2,3}$  is resolved by adding the rule  $C \rightarrow BC$ .

Finally the last recursion step that decides on the content of the root cell is shown in Figure 17. In this case the start variable is contained in the root of the pyramid.

Grammar :

$A \rightarrow BS \mid a$

$B \rightarrow CB \mid b$

$C \rightarrow BC \mid c$

$S \rightarrow BA \mid \mathbf{AC}$

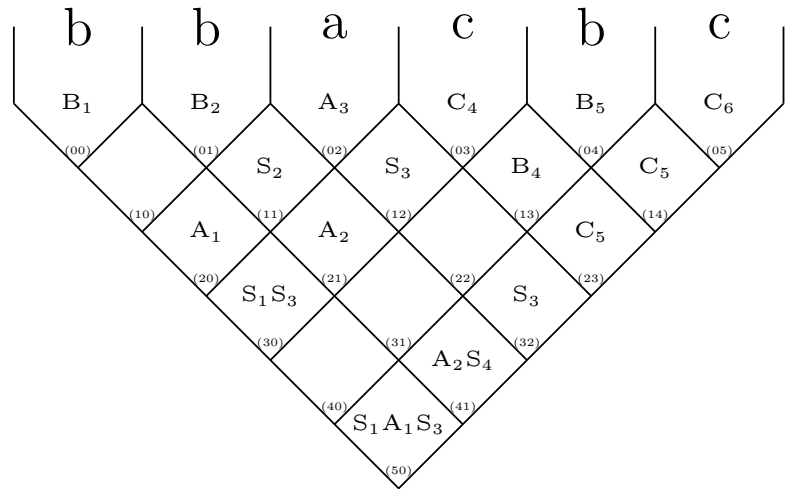


Figure 17: Illustration of Algorithm 7 SplitThenFillCYK part 6. The recursion step in  $Cell_{5,0}$  is resolved by adding the rule  $S \rightarrow AC$ .

## 2.6 Split Top-Down and fill Bottom-Up variant two

After defining an algorithm that uses the combination of the Bottom-Up approach and the Top-Down approach (Algorithm 7) a step further is to find an algorithm that only makes use of the Top-Down approach to see if an even better algorithm can be found. This algorithm again generates a predefined structure of the derivation tree Top-Downwards. Every time a node of the structure of the derivation tree has been decided on, a rule is immediately added to the grammar – therefore the name SplitThenFill, which is like "split for a node and then directly add a rule so that the node is then filled with at least one variable".

Note that the count of rules in the grammar is dependent on the count of nodes in the derivation tree and a terminal is distributed to only one variable. While resolving the last recursion step (Line 13) of Algorithm 10 SplitThenFillRec the start variable will be in the root of the pyramid that always leads to  $w \in L(G)$ .

<b>Algorithm 9:</b> SplitThenFill	
<b>Input:</b> Word $w \in \Sigma^*$	
<b>Output:</b> Set of rules $P$	
1	$P = \emptyset; \quad // \quad P \subseteq V \times (V^2 \cup \Sigma)$
2	$Sol = (P_{Sol}, v); \quad // \quad P_{Sol} \subseteq P$
3	$Sol = SplitThenFillRec(P, w, i_{max}, 0);$
4	<b>return</b> $P_{Sol};$
Line 2: $v$ can be any random element $v \in V$ .	

**Algorithm 10:** SplitThenFillRec**Input:**  $P_{in} \subseteq V \times (V^2 \cup \Sigma)$ ,  $w \in \Sigma^*$ ,  $i, j \in \mathbb{N}$ **Output:**  $(P, v)$ 

```

1  $P = P_{in};$ 
2 if  $i = 0$  then
3   if terminal  $w_j$  not distributed yet then
4     choose  $v \in V$  uniformly randomly;
5     return  $(P \cup \{(v, w_j)\}, v);$ 
6   end
7   return  $(P, v_{lhse});$ 
8 end
9 choose one  $m$  uniform randomly in  $[j + 1, j + i];$ 
10  $(P, v_l) = \text{SplitThenFillRec}(P, w, (m - j - 1), j);$ 
11  $(P, v_r) = \text{SplitThenFillRec}(P, w, (j + i - m), m);$ 
12 if  $i = i_{max}$  then
13   return  $(P \cup \{(S, v_l v_r)\}, S);$ 
14 end
15 choose  $v \in V$  uniformly randomly;
16 return  $(P \cup \{(v, v_l v_r)\}, v);$ 

```

Line 7: There is the rule  $v_{lhse} \rightarrow w_j$ , then  $v_{lhse}$  is the variable on the left side of the one rule that has the terminal  $w_j$  as its rhse.

**Example: Algorithm SplitThenFill**

According to this algorithm, only productions corresponding to the tree structure are added to the grammar. For illustration purposes, the pyramid is shown to reflect the immediate changes of the added rules to the pyramid in the Figures 18 to 23. Again the predefined derivation tree structure of Figure 11 is used.

At first the rule  $B \rightarrow b$  is added as shown in Figure 18.

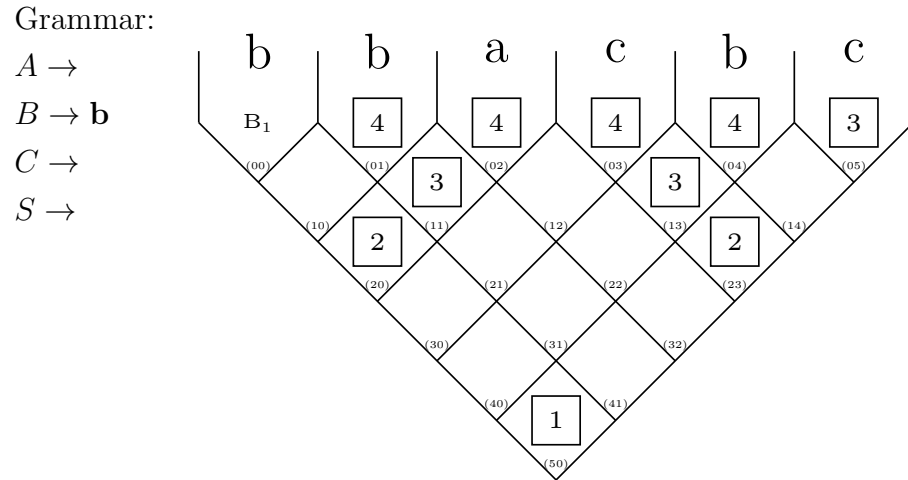


Figure 18: Illustration of Algorithm 9 SplitThenFill part 1. To resolve the recursion step that fills  $Cell_{0,0}$  the rule  $B \rightarrow b$  is added.

Next are the rules  $A \rightarrow a$  and  $S \rightarrow BA$  as seen in Figure 19. To resolve the recursion step that fills  $Cell_{0,1}$  no rule is added because a rule  $lhse \rightarrow b$  already exists. To fill  $Cell_{0,2}$  the rule  $A \rightarrow a$  and regarding  $Cell_{1,1}$  the rule  $S \rightarrow BA$  is added.

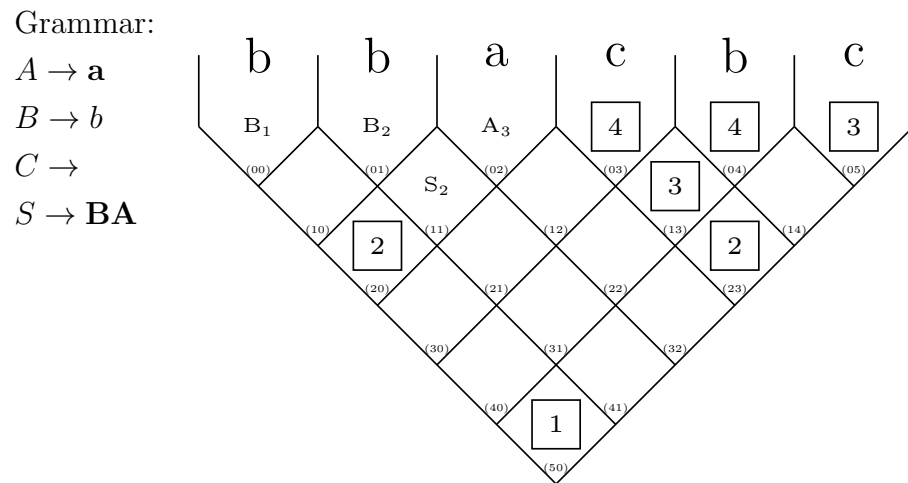


Figure 19: Illustration of Algorithm 9 SplitThenFill part 2. Adding of the rules  $A \rightarrow a$  and  $S \rightarrow BA$ .

To fill the  $Cell_{2,0}$  the rule  $C \rightarrow BS$  is added, see Figure 20:

Grammar:

$$A \rightarrow a$$

$$B \rightarrow b$$

$$C \rightarrow \mathbf{BS}$$

$$S \rightarrow BA$$

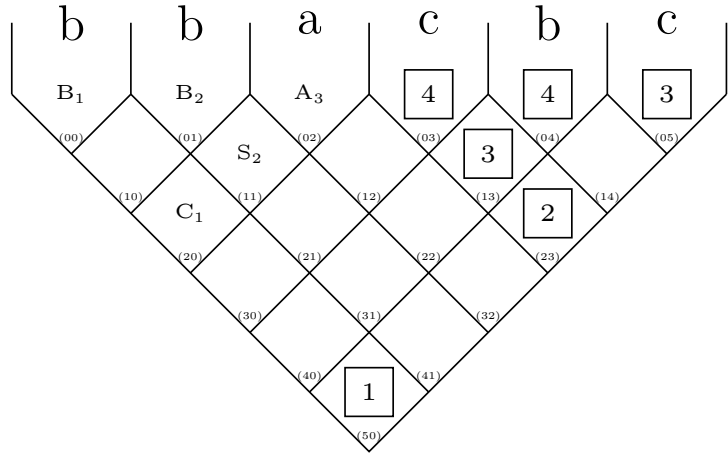


Figure 20: Illustration of Algorithm 9 SplitThenFill part 3. Adding of the rule  $C \rightarrow BS$  to fill the  $Cell_{2,0}$ .

Analogously the other cells are filled.  $Cell_{0,3}$  is responsible for the rule  $C \rightarrow c$ ,  $Cell_{0,4}$  does not cause a rule because again there already is the rule  $B \rightarrow b$ ,  $Cell_{1,3}$  contributes for the rule  $B \rightarrow CB$ ,  $Cell_{0,5}$  does not add a rule because of  $C \rightarrow c$  and to fill  $Cell_{2,3}$  the rule  $A \rightarrow BC$  is added.

Grammar:

$$A \rightarrow \mathbf{BC} \mid a$$

$$B \rightarrow \mathbf{CB} \mid b$$

$$C \rightarrow BS \mid c$$

$$S \rightarrow BA$$

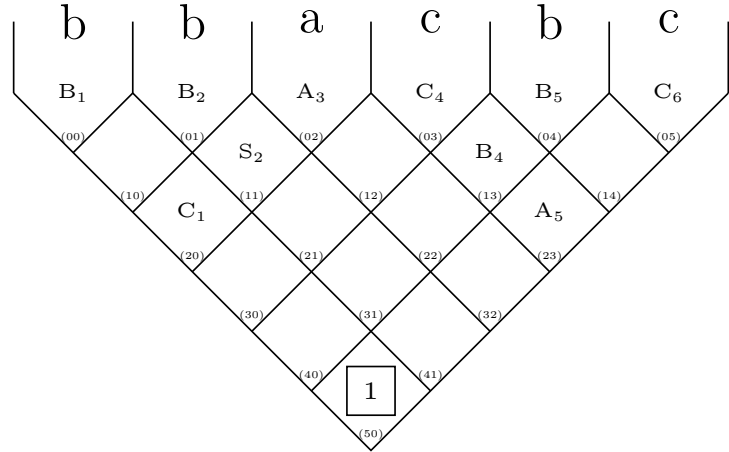


Figure 21: Illustration of Algorithm 9 SplitThenFill part 4. Adding of the rules  $C \rightarrow c$ ,  $B \rightarrow CB$  and  $A \rightarrow BC$ .

To fill the cell in the root a rule must be added that has the start variable as its *lhse* that guarantees  $w \in L(G)$ . Here the rule  $S \rightarrow CA$  is added.

Grammar:

$$A \rightarrow BC \mid a$$

$$B \rightarrow CB \mid b$$

$$C \rightarrow BS \mid c$$

$$S \rightarrow BA \mid \mathbf{CA}$$

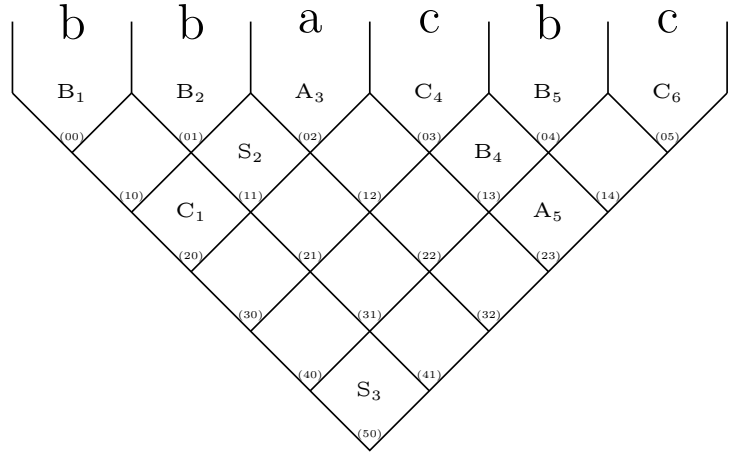


Figure 22: Illustration of Algorithm 9 SplitThenFill part 5. Adding of the rule  $S \rightarrow CA$ .

Finally a comparison of Figure 22 and Figure 23 shows the difference between parsing table after the last step of the algorithm and what the complete parsing table would look like. Additional variables are found in the cells  $Cell_{1,4}$ ,  $Cell_{2,3}$ ,  $Cell_{4,0}$  and  $Cell_{5,0}$ .

Grammar:

$$A \rightarrow BC \mid a$$

$$B \rightarrow CB \mid b$$

$$C \rightarrow BS \mid c$$

$$S \rightarrow BA \mid CA$$

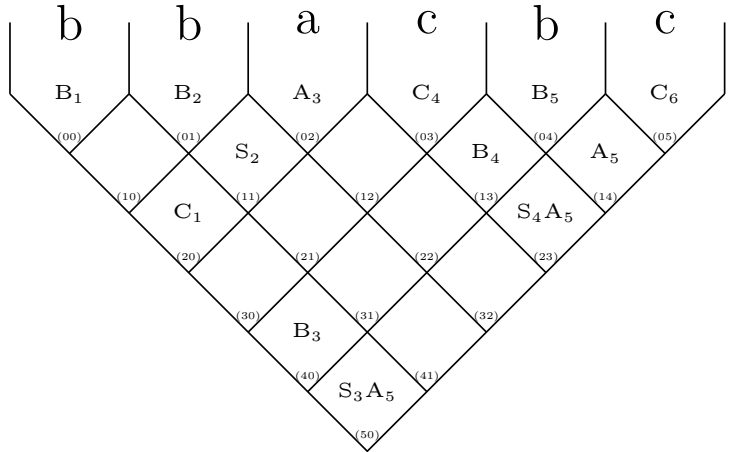


Figure 23: Illustration of Algorithm 9 SplitThenFill part 6. Comparison of the last step of the algorithm and the complete parsing table.

## 2.7 Evaluation of Algorithms

### 2.7.1 Success Rates

Until now different algorithms have been described that could be used in the application to create suitable exam exercises. But it is of interest to find out which algorithm performs the best and which algorithm should actually be used in the application to generate the exercises. Therefore a composite *Success Rate* is defined that measures the algorithms performance for the different requirements towards an exam exercise.

Here  $N \in \mathbb{N}$  is the sample size of all generated grammars while examining the algorithms. Before defining the overall Success Rate (*SR*) three other Success Rates set the basis for it.

**Success Rate Producibility:** A generated exercise contributes to the SR-Producibility if the CYK algorithm's output (Algorithm 1) is true or in other words  $w \in L(G)$ .  
 SR-Producibility =  $p/N$ , where  $p$  is the count of exercises that fulfil the requirement.

**Success Rate Cardinality-Rules:** A generated exercise contributes to the SR-Cardinality-Rules if the grammar has less than a certain count  $x$  of rules, i.e.  $|P| \leq x$  of the grammar  $G = (V, \Sigma, S, P)$ .  
 SR-Cardinality-Rules =  $cr/N$ , where  $cr$  is the count of exercises that fulfil the requirement.

**Success Rate Pyramid:** A generated exercise contributes to the SR-Pyramid if the following conditions are met:

1. At least one cell enforces to a correct cell combination – see the description of Algorithm 11 CheckforceCombinationPerCell for more information.
2. There are less than 100 variables in the entire pyramid.
3. There are less than 3 variables in each cell of the pyramid.

SR-Pyramid =  $p/N$ , where  $p$  is the count of exercises that fulfil the three requirements above.

For the sake of these checks (1., 2. and 3.) the definition of a *cell* in the *pyramid* is simplified as following:

$$Cell_{i,j} \subseteq \{(V, k) \mid k \in \mathbb{N}\} \longrightarrow Cell_{i,j} \subseteq V$$

For further illustration see Figure 24 on the next page.

**Example: Simplification of cell in a pyramid**

The pyramid with the simpler cells is shown right.

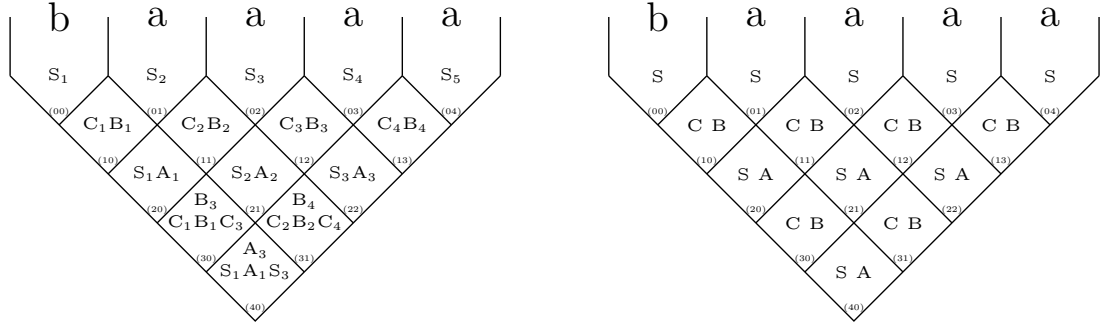


Figure 24: The simplification of cells in a pyramid.

**Description of the Algorithm CheckForceCombinationPerCell:**

The experience of professor Martens shows that usually most students easily find a pattern of how to fill the first two rows of the *pyramid* during the execution of the CYK algorithm but do more mistakes starting at row  $i \geq 2$ . Students often do not know exactly what cell combinations need to be considered while filling one specific cell of the pyramid. They simply take only the one top left cell and the one top right cell and try to find rules in the grammar that match the resulting compound variables. The approach of only finding patterns and not thoroughly understanding the algorithm is countered by Algorithm 11 CheckForceCombinationPerCell. Here a cell forces if it is possible to see if the student has clearly understood the algorithm and not only takes the next top left cell and the next top right cell.

<b>Algorithm 11:</b> CheckForceCombinationPerCell	
<b>Input:</b> $CellBottom, CellTopLeft, CellTopRight \subseteq V, P \subseteq V \times (V^2 \cup \Sigma)$	
<b>Output:</b> $true \iff  VarsForcing  > 0$	
1 $VarsForcing = \emptyset;$ // $VarsForcing \subseteq V$	
2 $VarComp = \{xy \mid x \in CellTopLeft \wedge y \in CellTopRight\};$	
3 <b>foreach</b> $v \in CellBottom$ <b>do</b>	
4	$Rhse = \{rhse \mid p \in P \wedge p = (v, rhse)\};$
5	<b>if</b> $Rhse \not\subseteq VarComp$ <b>then</b>
6	$VarsForcing = VarsForcing \cup v;$
7	<b>end</b>
8	<b>end</b>
9 <b>return</b> $ VarsForcing  > 0;$	
<p>Note: <math>CellBottom = Cell_{i,j}</math>, <math>CellTopLeft = Cell_{i-1,j}</math> and <math>CellTopRight = Cell_{i-1,j-1}</math>  Line 4: Get all rules of <math>P</math> that have <math>v</math> as the <math>lhse</math> and add their <math>rhse</math> to <math>Rhse</math>.  Line 5: If no <math>rhse \in Rhse</math> can be found in <math>VarComp</math>, then this variables forces, concluding that this cell as a hole forces.</p>	



**Example: Algorithm CheckForceCombinationPerCell**

In Figure 25, the variables in  $Cell_{2,0}$  and in  $Cell_{2,1}$  each force a right cell combination and in both cases  $VarComp = \{SS\}$ . The variable  $v = C$  does not have  $SS$  as one of its rhses and therefore the variable  $C$  forces.  $Cell_{3,0}$  does not force because  $VarComp = \{CC\}$  and the variable  $v = S$  has  $CC$  as its rhse. Note again, that cells with index  $i \leq 1$  can not force at all.

Grammar :

$C \rightarrow CS \mid a \mid b$

$S \rightarrow CC$

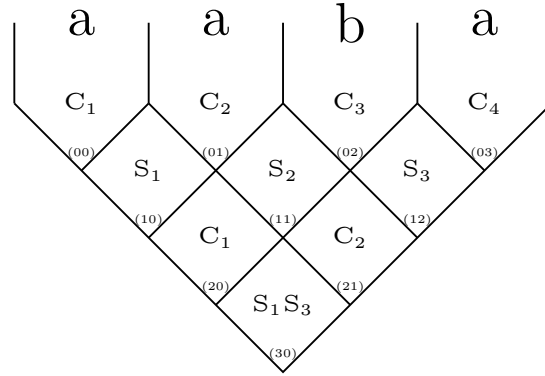


Figure 25: Application of Algorithm 11 CheckForceCombinationPerCell onto an entire pyramid.

The three success rates have been explained and finally the overall Success Rate can be specified.

**Success Rate:** A generated exercise contributes to the Success Rate (SR) if it contributes to the SR-Producibility, to the SR-Cardinality-Rules and to the SR-Pyramid at the same time.

It holds:  $SR = n/N$ , where  $n$  is the count of exercises.

### 2.7.2 Problem space exploration

Suitable ranges of parameters, to create exam exercises with, are:

- count of variables =  $[2; 8]$
- count of terminals =  $[2; 8]$
- size of word =  $[4; 11]$

This input parameter ranges are used during further comparison. Each calculated SR is based on a batch size  $N = 1024$ .

### 2.7.2.1 Comparison of the stopping criteria

As described in Chapter 2.1 Sub Modules, two different stopping criteria  $\textcircled{C}$  are used and it is of interest to know which one helps to generate more suitable exam exercises so that the better one can be used in a real world application. Therefore both variants of the stopping criteria are compared in Table 1 and Table 2.

Table 1 shows the average SRs over all configurations of parameters for each algorithm.

	MoreThanHalf	RootNotEmpty
DiceRollOnly	0.3%	<b>0.3%</b>
DiceRollVar1	0.5%	<b>1.3%</b>
DiceRollVar2	0.6%	<b>1.7%</b>
SplitThenFillCYK	1.4%	<b>1.4%</b>
SplitThenFill	8.4%	<b>8.4%</b>

Table 1: Average SRs over all the configurations of parameters for each algorithm.

In Table 2 the best possible SRs of each algorithm are compared.

	MoreThanHalf	RootNotEmpty
DiceRollOnly	17%	<b>17%</b>
DiceRollVar1	18%	<b>36%</b>
DiceRollVar2	20%	<b>38%</b>
SplitThenFillCYK	36%	<b>36%</b>
SplitThenFill	74%	<b>74%</b>

Table 2: Comparison of the best possible SRs of each algorithm. ( $N = 1024$ )

As seen in the two tables above (Table 1 and Table 2), the stopping criteria Root-NotEmpty performs better or at least equally good in every case. RootNotEmpty wins through and all further discussion in the following Chapter 2.7.2.2 is done with it.

### 2.7.2.2 Picking the best algorithm

Now that the choice of the stopping criteria has been decided in favour of Root-NotEmpty it is time to determine the one best algorithm.

As seen in Table 3 the Algorithm 9 SplitThenFill performs the best in all three sub success rates Producibility, Cardinality-Rules and Pyramid and therefore the algorithm also has the best average overall SR.

Algorithm	SR	Producibility	Cardinality-Rules	Pyramid			
					Force-Right	Vars-PerCell	VarsIn-Pyramid
DiceRollOnly	<b>0.3%</b>	5.2%	2.9%	12.1%	15.9%	24.9%	28.4%
BottomUpVar1	<b>1.3%</b>	7.7%	13.5%	11.9%	19.8%	20.8%	26.1%
BottomUpVar2	<b>1.7%</b>	10.9%	8.7%	13.5%	25.4%	16.6%	25.9%
SplitThenFillCYK	<b>1.4%</b>	4.3%	14.5%	14.7%	15.8%	27.6%	28.5%
SplitThenFill	<b>8.4%</b>	28.7%	17.4%	14.8%	18.3%	25.3%	27.3%

Table 3: More detailed comparison of the average SRs over all the configurations of parameters for each algorithm.

While looking at Table 4, that displays the best possible SRs, Algorithm 9 SplitThenFill again performs the best.

Algorithm	SR	Producibility	Cardinality-Rules	Pyramid			
					Force-Right	Vars-PerCell	VarsIn-Pyramid
DiceRollOnly	<b>17%</b>	21%	98%	47%	47%	100%	100%
BottomUpVar1	<b>36%</b>	60%	91%	65%	69%	100%	93%
BottomUpVar2	<b>38%</b>	56%	95%	69%	69%	100%	100%
SplitThenFillCYK	<b>36%</b>	51%	97%	71%	71%	100%	99%
SplitThenFill	<b>74%</b>	100%	100%	74%	74%	100%	100%

Table 4: More detailed comparison of the best possible SRs of each algorithm. (N = 1024)

The Algorithm 9 SplitThenFill wins through against the other four as it leads to the best SR on average and also has the best possible SR of all algorithms.

Therefore the Algorithm SplitThenFill is the best choice to use in an application.

## 3 GUI Tool: CYK Instances Generator

One of the goals of the thesis is to get a tool that assists in creating exam exercises. A Graphical User Interface (GUI) tool is preferred over a Command Line (CL) tool because of the ease of use and the better overview during the exercise creation.

### 3.1 Overview GUI

The developed tool consists of four major elements as marked in Figure 26.

1. Elementary input values can be given to the programm.
2. The status output of the programm is displayed.
3. Suitable exercises are calculated automatically and one can be selected.
4. The chosen exercise can be modified as wanted.

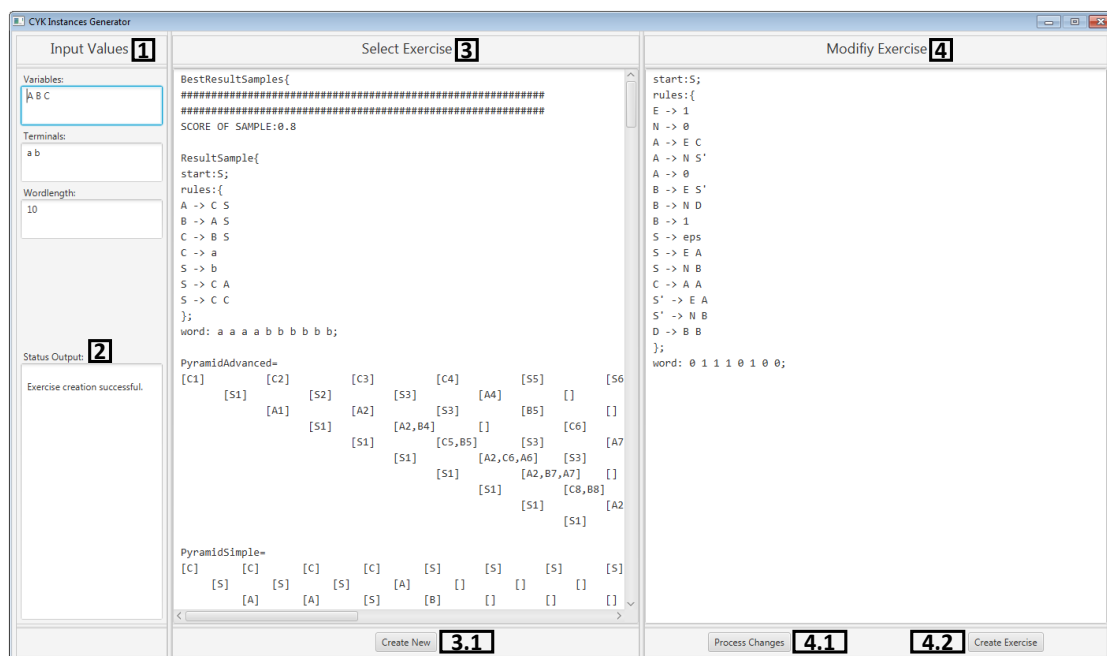


Figure 26: CYK Instances Generator.

Clicking the button 3.1 allows the creation of new suitable exercises with the given input values from area 1. Pressing button 4.1 processes the latest input given in area 4 to create a preview analogously to area 3 of how the created exercise would look like and finally button 4.2 creates the desired exercise. The exercise is created as a  $\text{\LaTeX}$ -code-file and a pdf-file. The  $\text{\LaTeX}$ -file is standalone compilable and allows further modification, whereas the pdf-file shows directly what the exercise looks like.

#### 3.1.1 Working with the program

The application structure contains only the executable "bachelor\_thesis\_cyk.jar"-file and the folder named "exercise". The mechanics of the application is mostly self explana-

tory. Just note that after clicking button 4.2 "Create Exercise" a new "exerciseLatex.tex"-file and the corresponding "exerciseLatex.pdf"-file will be generated within it and any files with the same name are overridden.

### 3.2 Exam Exercise

A exam *exercise* is a 4-tuple  $exercise = (grammar, word, parse table, derivation tree)$ .

#### Example: 4-tuple exercise

The pdf-file output of the tool looks similar to this:

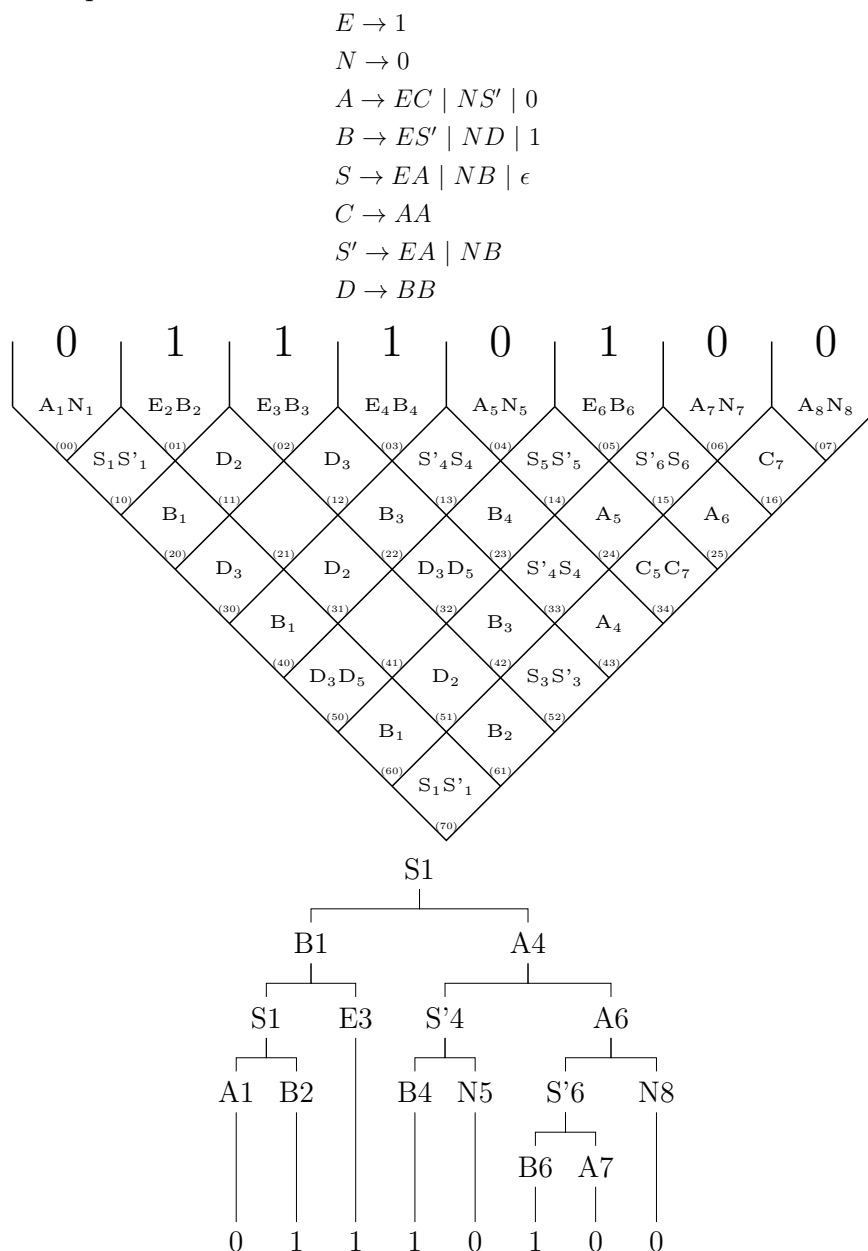


Figure 27: Example output for a *exercise*. Top: Context free grammar. Middle: A *pyramid* filled after the CYK algorithm. Bottom: Exemplary random derivation tree.

### 3.3 Scoring Model

Not every exercise is actually suitable for an exam. Therefore a scoring model is needed so that suitable exercise can be displayed in area 3 of the tool. Each exercise is given a score according to Table 5 and the parameters that influence the score are:

- `countRightCellCombinationsForced`, i.e. number of times a student is forced to make the right choice to fill the parsing table.
- `sumOfVarsInPyramid`, i.e. all variables in the pyramid.
- `countVarsPerCell`, i.e. maximum count of variables per cell.
- `sumOfRules`, i.e. all rules in the grammar.
- `countUniqueCells`, excluding row  $i = 0$ .

Parameter	Points					
	2	4	6	8	10	-100
<code>#cellCombinationsForced</code>	[0,10]	[11,20]	[21,30]	[41,50]	[31,40]	>50
<code>sumVarsInPyramid</code>	[0,10]	[11,20]	[21,30]	[41,50]	[31,40]	>50
<code>#VarsPerCell</code>	[5,5]	[4,4]	[1,1]	[3,3]	[2,2]	>5
<code>sumOfRules</code>	[1,2]	[3,4]	[5,6]	[9,10]	[7,8]	>10
<code>countUniqueCells</code>	[3,3]	[4,4]	[5,5]	[6,6]	[7,7]	$\leq 2$

Table 5: Scoring of the different parameter values.

The score of each *exercise* is normalized to the maximum possible points so the maximum score is 1.0:  $score = (\#Parameter \cdot 10)^{-1} \cdot \sum_{parameter} points$ .

One negative score is already sufficient to avoid examples in area 3 with undesired properties. One negative score is already sufficient that the overall score of the exercise is negative.

### 3.4 Parsing input with ANTLR

In area 4 of the application a context free grammar is given as input as seen in Figure 28 on the left. This input is parsed with ANTLR <sup>4</sup> because it allows a clear separation between the language definition and the Java code.

The first step here is the tokenization of the input in token as seen in Figure 28 on the right. After that with the help of the Grammar as seen in Figure 29 an abstract syntax tree is generated out of which Java source code is generated.

start:S;	
rules:{	
E -> 1	START: ('start');
N -> 0	RULES: ('rules');
A -> E C	ARROW: ('->');
A -> N S'	WORD: ('word');
A -> 0	
B -> E S'	UPPERCASE: ('A'..'Z');
B -> N D	LOWER_CASE_OR_NUM: ('a'..'z'   '0'..'9');
B -> 1	
S -> eps	OPEN_BRACE: '(';
S -> E A	CLOSE_BRACE: ')';
S -> N B	OPEN_BRACE_CURLY: '{';
C -> A A	CLOSE_BRACE_CURLY: '}';
S' -> E A	
S' -> N B	SEMICOLON : ';';
D -> B B	COLON: (':');
};	WHITE_SPACE: ' '   '\t';
word: 0 1 1 1 0 1 0 0;	NEWLINE: '\n';
	SPECIALSYMBOL: ('\');

Figure 28: Left: Input grammar example of the application. Right: Formal definition of the used ANTLR grammar tokens to parse the input grammar.

<sup>4</sup>ANTLR works with LL(k) grammars, which means that each derivation step can be distinctly identified through the next k tokens.

```

grammar Exercise;

exerciseDefinition: grammarDefinition NEWLINE
                    wordDefinition NEWLINE?;

grammarDefinition: NEWLINE* WHITE_SPACE* varStart WHITE_SPACE* NEWLINE
                    rules;

varStart: START COLON WHITE_SPACE* nonTerminal SEMICOLON;

rules: RULES COLON WHITE_SPACE* OPEN_BRACE_CURLY NEWLINE
      (singleRule NEWLINE)+
      CLOSE_BRACE_CURLY SEMICOLON;

singleRule: WHITE_SPACE* nonTerminal // A
            WHITE_SPACE* ARROW WHITE_SPACE* // ->
            terminal WHITE_SPACE* // a
            |
            WHITE_SPACE* nonTerminal // A
            WHITE_SPACE* ARROW WHITE_SPACE* // ->
            nonTerminal WHITE_SPACE+ nonTerminal WHITE_SPACE*;

wordDefinition: WORD COLON WHITE_SPACE* terminals WHITE_SPACE* SEMICOLON;

terminals: terminal
          |
          terminal WHITE_SPACE terminals;

nonTerminal: UPPERCASE+ SPECIALSYMBOL?;
terminal: LOWER_CASE_OR_NUM+;

```

Figure 29: Formal definition of the used ANTLR grammar rules.

### 3.5 Other Matters

Lastly, just some general information about the implementation are given here.

Technologies that have been used for programming are Github <sup>5</sup> with Sourcetree <sup>6</sup> for version control, Maven <sup>7</sup> for build management, IntelliJ <sup>8</sup> as the IDE, ANTLR <sup>9</sup> with ANTLRWorks for parsing input and JavaFX Scene Builder <sup>10</sup> to create the GUI.

Important used frameworks are: JUnit <sup>11</sup> for testing and Project Lombok <sup>12</sup> to greatly reduce boilerplate code.

Altogether around 7100 lines of code have been written, of which 5400 are pure java code lines, 900 are comment lines and 800 are blank lines.

<sup>5</sup>Github: <https://github.com/>

<sup>6</sup>Sourcetree: <https://www.sourcetreeapp.com/>

<sup>7</sup>Maven: <https://maven.apache.org/>

<sup>8</sup>IntelliJ: <https://www.jetbrains.com/idea/>

<sup>9</sup>ANTLR: <http://www.antlr.org/>

<sup>10</sup>JavaFX Scene Builder: <http://gluonhq.com/products/scene-builder/>

<sup>11</sup>JUnit: <http://junit.org/junit4/>

<sup>12</sup>Project Lombok: <https://projectlombok.org/features/all>



## References

- [1] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2. Aufl. edition, 2012.
- [2] Itiroo Sakai. *Syntax in universal translation*. Her Majesty's Stationery Office, London, 1962.
- [3] John Cocke, Jacob T. Schwartz. *Programming languages and their compilers. Preliminary notes*. Courant Institute of Mathematical Sciences of New York University, New York, 1970.
- [4] Tadao Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. 1966.
- [5] Daniel Younger. Recognition and parsing of context-free languages in time  $n^3$ . *INFORMATION AND CONTROL*, 10(2):189–208, 1967.
- [6] Dirk Hoffmann. *Theoretische Informatik*. Hanser, Carl, München, 1., neu bearbeitete auflage edition, 2015.

## Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

*Unterschrift :*

*Ort, Datum :*

