# Assignment 6
# Context-free Languages

Josefin Ulfenborg
940806-5960
yunalescca@gmail.com

2017-05-xx

# 1 Task 1

**a)** The only two nullable characters we have are A and C. After eliminating them the grammar will be updated to the following:

$$S \to aAbB \mid abB \mid cCdD \mid cdD \mid fFgG$$

$$A \to E \mid aA \mid a$$

$$B \to bC \mid b \mid bb \mid CC \mid C$$

$$C \to cC \mid c \mid D$$

$$D \to dD \mid d \mid E$$

$$E \to Ee \mid eE$$

$$G \to f \mid g$$

**b)** The unit productions in the grammar are A ($A \to E$), B ($B \to C$), C ($C \to D$) and D ($D \to E$). After eliminating these the grammar looks like this:

$$S \to aAbB \mid abB \mid cCdD \mid cdD \mid fFgG$$

$$A \to Ee \mid eE \mid aA \mid a$$

$$B \to bC \mid b \mid bb \mid CC \mid cC \mid c \mid dD \mid d \mid Ee \mid eE$$

$$C \to cC \mid c \mid dD \mid d \mid Ee \mid eE$$

$$D \to dD \mid d \mid Ee \mid eE$$

$$E \to Ee \mid eE$$

$$G \to f \mid g$$

**c)** The useless symbols are either the non-reachable or the ones that do not produce anything to the language. The first thing to do is locate the generating states and eliminate those who are not.

The base case is $\{a, b, c, d, e, f, g\}$, and all of these terminals are generating since they generate themselves. Then, from the inductive step, we get that S, A, B, C, D, G are generating. That is, E and F are not generating. E is not generating because from E we cannot derive a string with only terminals, and F is not generating because we do not have any productions for F. In order to eliminate them I simply remove the symbols and the productions. Hence the grammar will now look like this:

$$S \to aAbB \mid abB \mid cCdD \mid cdD$$

$$A \to aA \mid a$$

$$B \to bC \mid b \mid bb \mid CC \mid cC \mid c \mid dD \mid d$$
$$C \to cC \mid c \mid dD \mid d$$
$$D \to dD \mid d$$
$$G \to f \mid g$$

The second part is now to eliminate the non-reachable symbols. The base case is S, which is reachable. From S we can deduce that a, b, c, d, A, B, C and D are reachable. Since I previously removed the production $A \to fFgG$, f, g and G are no longer reachable and so I remove these from the grammar. Now it looks like this:

$$S \to aAbB \mid abB \mid cCdD \mid cdD$$
$$A \to aA \mid a$$
$$B \to bC \mid b \mid bb \mid CC \mid cC \mid c \mid dD \mid d$$
$$C \to cC \mid c \mid dD \mid d$$
$$D \to dD \mid d$$

**d)** Now finally I can give the Chomsky Normal Form (CNF) of the grammar. A grammar in CNF has productions only of the form $X \to t$ or $X \to YZ$ (exactly two symbols). The algorithm gives me the following, final grammar, where I have put the original symbols at top of the list, and the newly introduced symbols thereafter:

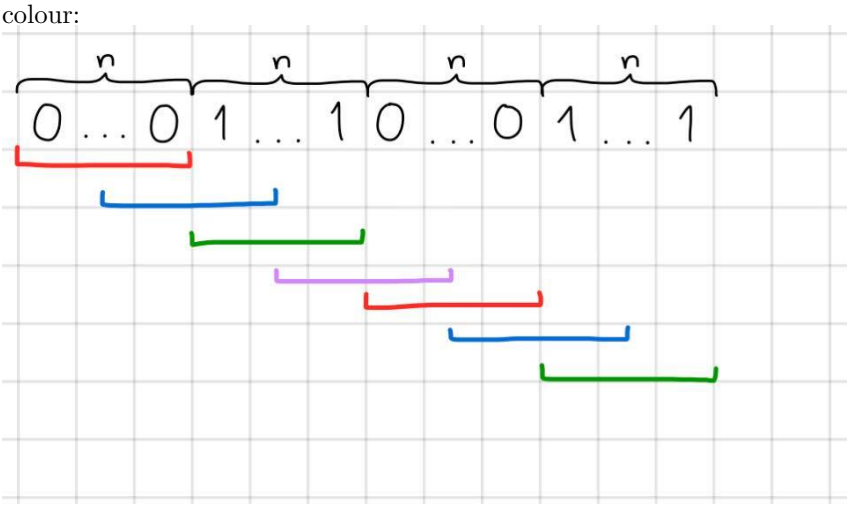
$$S \to X_A A_1 \mid X_A B_1 \mid X_c C_1 \mid X_c D_1$$
$$A \to X_A A \mid a$$
$$B \to X_B C \mid b \mid X_B X_B \mid CC \mid X_C C \mid c \mid X_D D \mid d$$
$$C \to X_C C \mid c \mid X_D D \mid d$$
$$D \to X_D D \mid d$$
$$A_1 \to A B_1$$
$$B_1 \to X_B B$$
$$C_1 \to C D_1$$
$$D_1 \to X_D D$$
$$X_A \to a$$
$$X_B \to b$$
$$X_C \to c$$
$$X_D \to d$$

## 2 Task 2

Given language: $\mathcal{L} = \{0^i 1^j 0^i 1^j \mid i+j > 0\}$. That is i or j is $\neq 0$, so the smallest word in the language will be 00 or 11. if both i and j are $\neq 0$, then the smallest word is 0101. The empty word, 011 or 001 should not be possible, however. The language states that we must have an equal amount of 0's in both "halves" of the word, and an equal amount of 1's in both "halves".

Now for the pumping lemma, where I will prove this language is *not* CF.

1. The first step is that I will assume that my language $\mathcal{L}$ *is* context-free, so now I can perform the pumping lemma on it.
2. Then by the pumping lemma we will get a constant $n > 0$
3. The next step is to pick a word w that is in the language of $\mathcal{L}$. I will pick the word $w = 0^n 1^n 0^n 1^n$. Clearly we have that $|w| \geq n$
4. Now, by the pumping lemma we know that $\exists_{x,u,y,v,z}.\ w = xuyvz$, such that
   - $|uyv| \leq n$
   - $uv \neq \epsilon$, that is, either u or v is not empty
   - $\forall k \geq 0.\ xu^k yv^k z \in \mathcal{L}$
5. Now, because of the word I have chosen, and because I know that $|uyv| \leq n$, there are certain restrictions on what the sub word $uyv$ can consist of. I have made a simple drawing displaying the different possibilities and splittings of the sub word. I have marked similar splittings in the same colour:




This means I have four different cases, and I will prove all of them in order to show, by contradiction that $\mathcal{L}$ is not CF. One thing to note is that the parts in blue and purple can be proven at the same time.

First I will introduce two variables $q, r > 0$. Then, on each possibility of a sub word, I will suggest a value for k such that $xu^k yv^k z \notin \mathcal{L}$.

i (coloured red) $uyv = 0^q$

Here $x$ may be empty, or consist of some (but only) 0's and $z$ will be the remaining word after the (at most) n 0's.

For this sub word it is enough to pick k = 0, leaving us with $w' = xyz$. And since either u or v is non-empty, one of them has to consist of at least one 0, and then we will have fewer 0's in the beginning of the word than what we had before. That is, the number of 0's in the first half of the word, is less than the number of 0's in the second half of the word, so, $w' \notin \mathcal{L}$

ii (coloured blue) $uyv = 0^q 1^r$

Same condition for $x$ as above. $z$ will be the remaining word after $uyv$ and $uyv$ will consist of at least one 0 *and* at least one 1.

Moreover, choosing k = 0 shows to be enough here as well. By removing $u$ and $v$ from the word, and (again) since either $u$ or $v$ is non-empty, we either remove at least one 0 or at least one 1, or both of them. This gives us one of the following:

- We now have fewer 0's in the first half of the word than what we had before, so thew new word cannot be in the language.
- We now have fewer 1's in the first half of the word, so the word cannot be in the language here either.
- We have fewer 0's and 1's in the first half, so again this word cannot be in the language

Of course, I can switch 'first half' for 'second half' if that's the part of the word I'm talking about.

iii (coloured green) $uyv = 1^q$

The same tactic goes here as for the first example, so choosing k = 0 is enough. Then the first block of 1's are fewer than before and so this new word cannot be in the language.

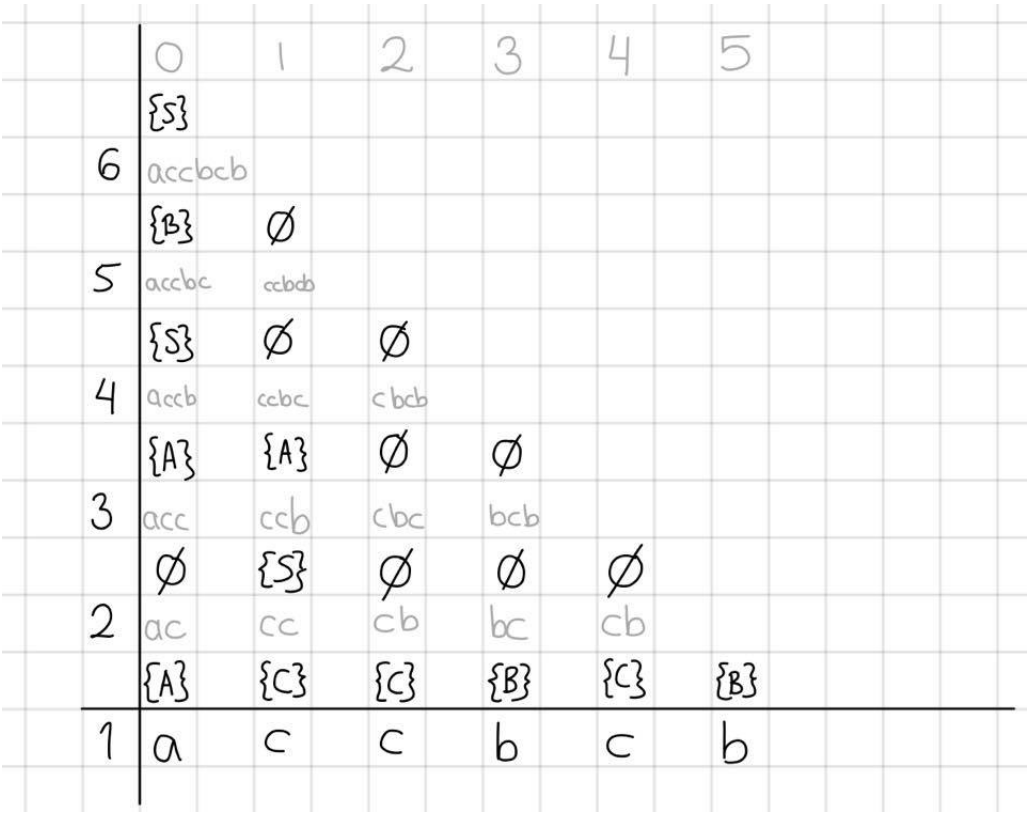
iv (coloured purple) $uyv = 1^q 0^r$

As I mentioned before, by proving ii), I have also proven iv), so the same solution goes here.

6. In all cases I found a k such that $xu^k yv^k z \notin \mathcal{L}$, so by contradiction, it results in that our assumption that $\mathcal{L}$ is CF is false. L is not a CFL.

## 3 Task 3

Here is the table I retrieved after performing the CYK algorithm on the grammar:



| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| | {S} | | | | | |
| 6 | accbcb | | | | | |
| | {B} | ∅ | | | | |
| 5 | accbc | ccbcb | | | | |
| | {S} | ∅ | ∅ | | | |
| 4 | accb | ccbc | cbcb | | | |
| | {A} | {A} | ∅ | ∅ | | |
| 3 | acc | ccb | cbc | bcb | | |
| | ∅ | {S} | ∅ | ∅ | ∅ | |
| 2 | ac | cc | cb | bc | cb | |
| | {A} | {C} | {C} | {B} | {C} | {B} |
| 1 | a | c | c | b | c | b |

My conclusion from this is that *yes*, the word *accbcb* is in the grammar, and this because on the last row the starting symbol S is in the set.