8. Pumping Lemmas for Context-Free Languages

Goal: Introduce pumping lemmas for context-free languages that allow us to show that a given language is not context-free.

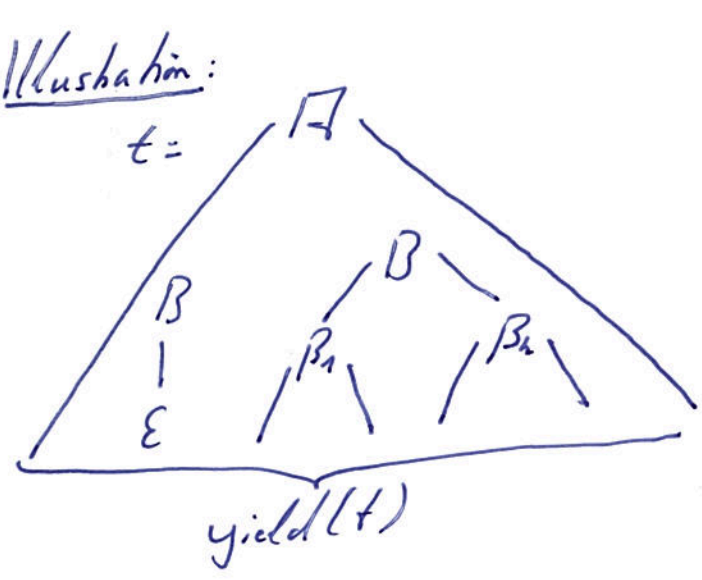
8.1 Parse Trees

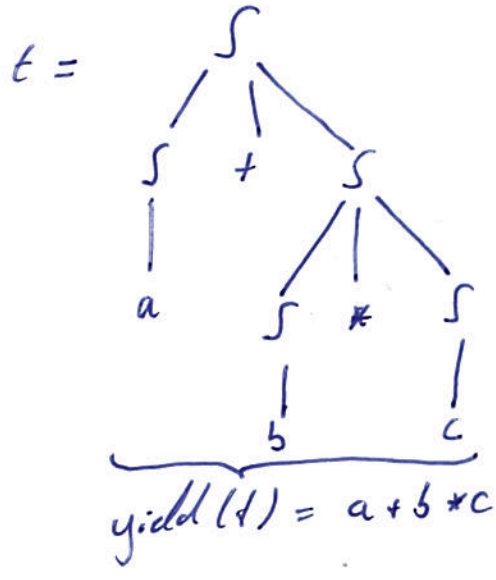
Definition:

- A parse tree from a non-terminal B in $G = (N, \Sigma, P, S)$ is a tree with the following properties:
  - Every node is labelled with a symbol from $N \cup \Sigma \cup \{\varepsilon\}$.
    The root is labelled by B.
    Every inner node is labelled by a symbol from N.
  - If B is an inner node with k children labelled $\beta_1, \dots, \beta_k$ (from left to right), then
    - $\beta_1, \dots, \beta_k \in N \cup \Sigma$ and
      $B \to \beta_1 \dots \beta_k \in P$

    or
    - $k = 1$, $\beta_1 = \varepsilon$, and $B \to \varepsilon \in P$.
- The yield of a parse tree t, denoted by $\text{yield}(t) \in (N \cup \Sigma)^*$, is the concatenation of the leaves from left to right.

Illustration:





Definition:

The parse tree associated with a derivation

$$\Pi = \alpha_0 \Rightarrow \dots \Rightarrow \alpha_n$$

is defined by induction on the length of the derivation:

$n=0$: For the empty derivation from A to A, we obtain the parse tree with a single node (root and leaf) A

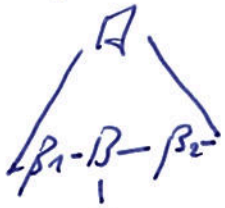
$n \to n+1$: Consider

$$\underbrace{A}_{\alpha_0} \to \dots \to \underbrace{\beta_1 B \beta_2}_{\alpha_n} \to \underbrace{\beta_1 \gamma \beta_2}_{\alpha_{n+1}}.$$

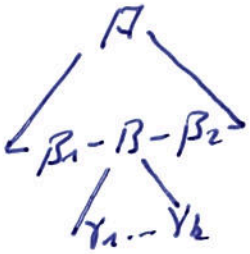
Let t be the parse tree for $A \to \dots \to \beta_1 B \beta_2$.

We extend t depending on $B \to \gamma$.

↳ If $\gamma = \varepsilon$, we obtain



↳ If $\gamma = \gamma_1 \dots \gamma_k$ with $\gamma_i \in N \cup \Sigma$, we get



Theorem:

Let $G = (N, \Sigma, P, S)$, $A \in N$, and $\alpha \in (N \cup \Sigma)^*$.

(1) $A \Rightarrow^* \alpha$ iff there is a parse tree t from A with $yield(t) = \alpha$.

(2) For every derivation $A \Rightarrow^* \alpha$, there is a unique parse tree (the one associated with it).

(3) The same parse tree may be associated with several derivations, but with only one left-derivation and only one right-derivation

8.2 The Pumping Lemma

Goal: Introduce the classic pumping lemma for CFLs.

Theorem (Pumping lemma, Bar-Hillel, Perles, Shamir '61):

Let L be a CFL.

There is a constant p_L so that for all $z \in L$ with $|z| \geq p_L$

there is a decomposition

$$z = uvwxy$$

satisfying

(1) $|vx| \geq 1$

(2) $|vwx| \leq p_L$

(3) for all $i \in \mathbb{N}$, $uv^i wx^i y \in L$.

Proof:

Let G be a context-free grammar in Chomsky normal form generating $L \setminus \{\varepsilon\}$.

Let k be the number of non-terminals in G.

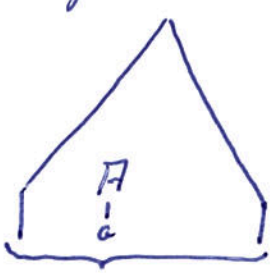
We define

$$p_L := 2^k.$$

Consider a word $z \in \Sigma^*$ with $|z| \geq p_L$.

- The derivation tree of z is – due to the Chomsky normal form – a binary tree, up to the last step $A \to a$:

$t =$

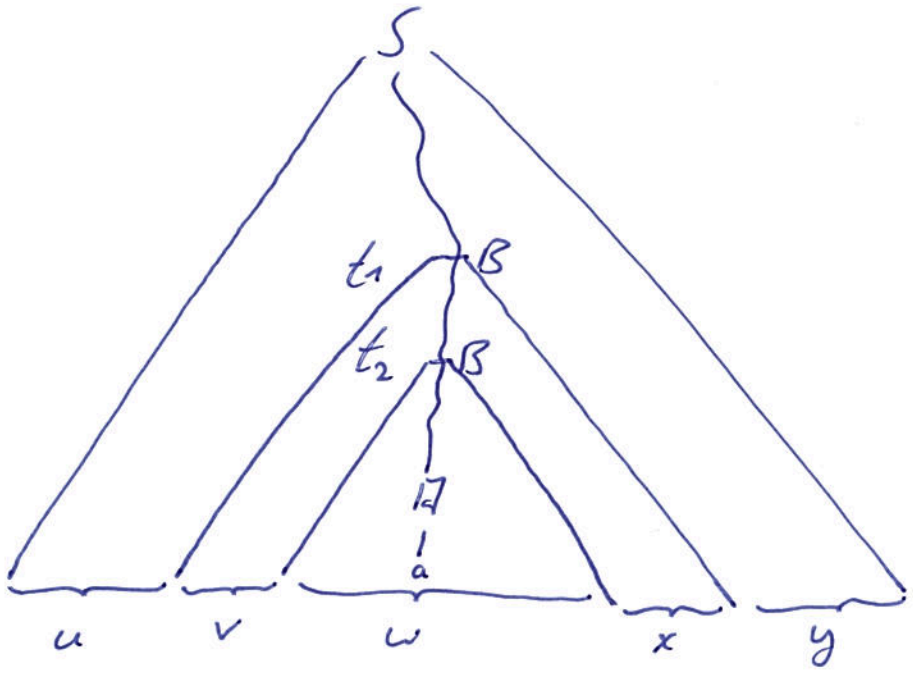


$\text{yield}(t) = z$ with $|z| \geq p_L$.

- Since the tree has at least 2^k leaves and the last step is $A \to a$, it has height at least $k+1$.

  Indeed, a full binary tree of height n has 2^n leaves.

- Fix a longest path (which has $\geq k+2$ nodes).
  There are $\geq k+1$ non-terminals on this path.
  But the grammar only has k non-terminals.
  Hence, by the pigeon hole principle, one non-terminal has to repeat.
  We consider the first repetition from the leaves.



- The upper B is at most $k+1$ steps from the leaves.
  This means the subtree t_1 rooted at the upper B has a yield of length $\leq 2^k$.
  Let t_2 be the subtree rooted at the lower B.
  Let w be the yield of t_2.
  We now have

  $$\text{yield}(t_1) = v\,w\,x \quad \text{for suitable } v, x \in \Sigma^* \text{ with } |vwx| \leq 2^k = p_L$$

- We have $v \neq \varepsilon$ or $x \neq \varepsilon$.
  To see this, note that the first production applied to the upper B is of the form $B \to CD$.
  Now t_2 is completely inside the tree rooted at C or inside the tree rooted at D.
  Hence, x is (at least partly) derived from D or v is (at least partly) derived from C.
  Therefore, x or v is not ε.
- Finally, we note that
  $$B \Rightarrow^* vBx \quad \text{and} \quad B \Rightarrow^* w.$$
  Hence, $B \Rightarrow^* v^i w x^i$ for all $i \in \mathbb{N}$. □

Example:
$L = \{a^n b^n c^n \mid n \in \mathbb{N}\}$ is not context-free.

Proof:
Assume L was context-free with pumping constant p_L.
Consider $a^{p_L} b^{p_L} c^{p_L}$.
We decompose the word into $uvwxy$ so as to satisfy the conditions of the pumping lemma.
We ask ourselves where v and x, the words that get pumped, could lie.
Because $|vwx| \leq p_L$, it is impossible for vx to contain a's and c's.

- If vx consists of a's only, then $uv^0wx^0y = uwy$ has p_L many b's and p_L many c's, but $< p_L$ many a's, since $|vx| \geq 1$. ↯

- If vx consists of b's only, of c's only,
  of a's and b's only, and of b's and c's only,
  the contradiction is derived similarly.

  Hence, L is not context-free. □

8.3 Ogden's Lemma

Goal: · Prove a stronger pumping lemma for context-free languages.
· The usual pumping lemma sometimes does not apply
(see the example below).

Theorem (Ogden '68):

Let L be a context-free language.
There is a constant p_L so that for all $z \in L$
where we mark $\geq p_L$ positions as distinguished
there is a decomposition

$$z = uvwxy$$

with
(1) v and x together have at least one distinguished position
(2) vwx has at most p_L distinguished positions
(3) for all $i \in \mathbb{N}$, $uv^i wx^i y \in L$.

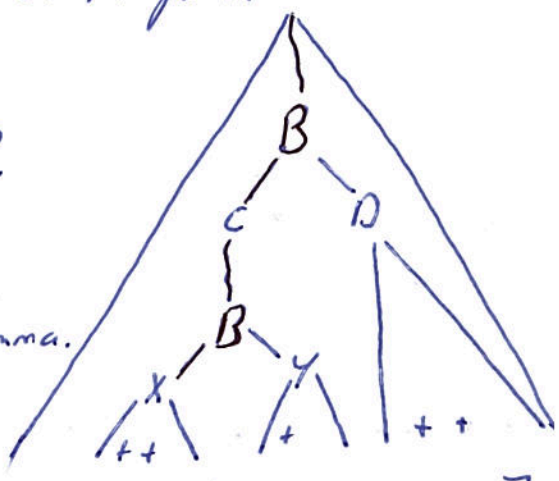
Proof:
- Let G be in Chomsky normal form generating $L \setminus \{\varepsilon\}$.
  Let G have k non-terminals and define $p_L := 2^k + 1$.
  Let z be a word with $\geq p_L$ positions marked.
- Our first goal is to construct a path
  in the parse tree for z's derivation.

Since we have to care about the distinguished positions,
we cannot just pick a longest path.
Moreover, we cannot pick arbitrary repeating non-terminals.
Instead, we have to consider branch points:
non-terminals where both children have distinguished descendants.

- We construct the path inductively:
  - ↳ The root is on the path.
  - ↳ Suppose r is the last node on the path.
    - ↳ If r is a leaf, the construction ends.
    - ↳ If r has only one child w.l.o.g. with distinguished descendants, add that child to the path and continue from there.
    - ↳ If both children of r have distinguished descendants, call r a branch point and continue from the child with the larger number of distinguished descendants (arbitrary if equal).

- Each branch point has at least half as many distinguished descendants as the previous branch point.
Since there are strictly more than 2^k (namely $\geq 2^k+1$) distinguished positions in z,
there are at least $k+1$ branch points on the path.

- Among the last $k+1$ branch points
there are two with the same non-terminal B.
Select the two closest to the leaves
and argue like in the previous pumping lemma.



□

Remark (Conservative extension):

If we apply Ogden's lemma to a word where all positions are marked, we obtain the classic pumping lemma.

Example (Application of Ogden's lemma):

$L = \{a^i b^j c^k \mid i \neq j,\ j \neq k,\ \text{and } k \neq i\}$ is not context-free.

Proof:

Suppose L was context-free.

Let p_L be the constant in Ogden's lemma.

We consider the word

$$z = a^{p_L} b^{p_L + p_L!} c^{p_L + 2p_L!} \in L.$$

Let the positions of the a's be distinguished.

Let

$$z = uvwxy$$

be a decomposition that satisfies the conditions in Ogden's lemma.

↳ If v or x contains two different symbols,

then $uv^2wx^2y \notin L$.

For example, if $v \in a^+b^+$, then vv has a b followed by an a.

↳ Now at least one of v and x must contain a's,

since only a's are distinguished positions.

So if $x \in b^*$ or $x \in c^*$, then $v \in a^+$.

Moreover, if $x \in a^+$ then $v \in a^*$.

↳ We focus on the situation where $x \in b^*$ and $v \in a^+$.

The remaining cases are similar.

Note that $1 \leq |v| \leq p_L$ since there are at most p_L-many a's.
This means $|v|$ divides $p_L!$ $(= p_L \cdot (p_L - 1) \cdot (p_L - 2) \cdot \ldots \cdot 2 \cdot 1)$.
Let q be such that $|v| \cdot q = p_L!$.

Then
$$z' = u\, v^{2q+1}\, w\, x^{2q+1}\, y \in L.$$

But z' has
$$\underbrace{p_L - |v|}_{a^{p_L - |v|}} + (2q+1) \cdot |v| = p_L + 2 \cdot \underbrace{q \cdot |v|}_{p_L!} = p_L + 2p_L! \text{ many a's.}$$

However, since v and x have no c's,
z' also has $p_L + 2p_L!$ many c's. $\lightning$ $z' \notin L$.

A similar contradiction occurs if $x \in a^+$ or $x \in c^*$.
Therefore, L is no context-free language. □