Automata Theory and Formal Languages

Summer Term 2026

12$^{\text{th}}$ Lecture

The Pumping Lemma for Context-Free Languages

Prof. Dr. Matthias Mnich *Institute for Algorithms and Complexity*

Objectives

We study a procedure to refute the context-freeness of certain languages.

Introducing the pumping lemma for CFLs

Recall the pumping lemma for regular languages:
Any regular language satisfies the *pumping property* for regular languages:
One can pump up words $x \in L$ at **one position** and $x' = uv^i w$ remains in L for all i.

Introducing the pumping lemma for CFLs

Recall the pumping lemma for regular languages:
Any regular language satisfies the *pumping property* for regular languages:
One can pump up words $x \in L$ at **one position** and $x' = uv^iw$ remains in L for all i.

Analogously we will establish for context-free languages:
Any context-free language satisfies the *pumping property* for context-free languages:
One can pump up words $x \in L$ at **two positions simultaneously** and $x' = uv^iwx^iy$ remains in L for all i.

Introducing the pumping lemma for CFLs

Recall the pumping lemma for regular languages:
Any regular language satisfies the *pumping property* for regular languages:
One can pump up words $x \in L$ at **one position** and $x' = uv^i w$ remains in L for all i.

Analogously we will establish for context-free languages:
Any context-free language satisfies the *pumping property* for context-free languages:
One can pump up words $x \in L$ at **two positions simultaneously** and $x' = uv^i wx^i y$ remains in L for all i.

Chiefly used to establish non-context-freeness:
Some language L violates the pumping property for CFLs.
$\Rightarrow$ L is not context-free.

Pumping lemma for context-free languages

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$,
- $|vwx| \leq n$,
- $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$,
- $|vwx| \leq n$,
- $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

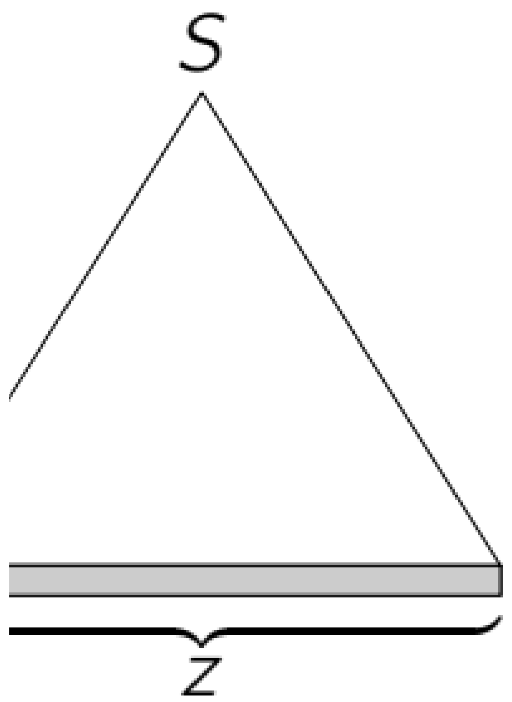
- Let $G = (V, \Sigma, P, S)$ be a context-free grammar in Chomsky normal form with $L(G) = L \setminus \{\varepsilon\}$.
- Consider the parse tree T_z of any word $z \in L$ with $|z| \geq 2^{|V|} =: n$.

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

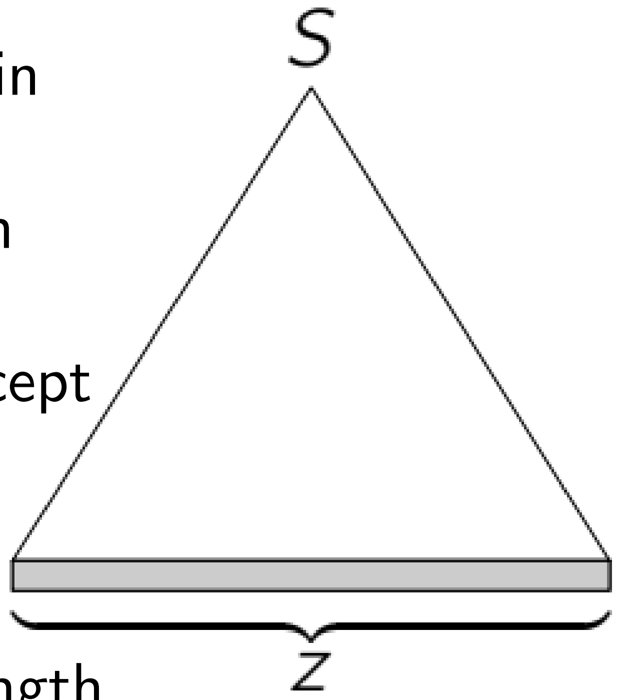
- Let $G = (V, \Sigma, P, S)$ be a context-free grammar in Chomsky normal form with $L(G) = L \setminus \{\varepsilon\}$.
- Consider the parse tree T_z of any word $z \in L$ with $|z| \geq 2^{|V|} =: n$.



Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.



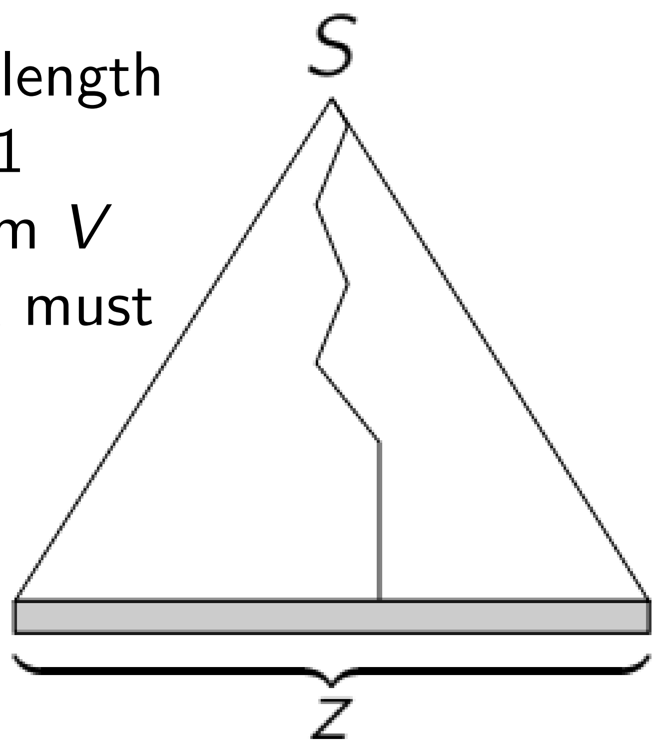
- Let $G = (V, \Sigma, P, S)$ be a context-free grammar in Chomsky normal form with $L(G) = L \setminus \{\varepsilon\}$.
- Consider the parse tree T_z of any word $z \in L$ with $|z| \geq 2^{|V|} =: n$.
- As G is in Chomsky normal form, T_z is binary except for maybe the bottom (leave) layer which applies productions rules $A \rightarrow a$.
- $\rightarrow$ T_z without bottom layer has $|z| \geq 2^{|V|}$ leaves
- Hence, a longest path π from root to leave has length at least $|V|$, where π consists of at least $|V| + 1$ nodes each of which is labeled by a variable from V

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

- ...
- Hence, a longest path π from root to leave has length at least $|V|$, where π consists of at least $|V| + 1$ nodes each of which is labeled by a variable from V
- As there are only $|V|$ variables, some variable A must appear twice on π

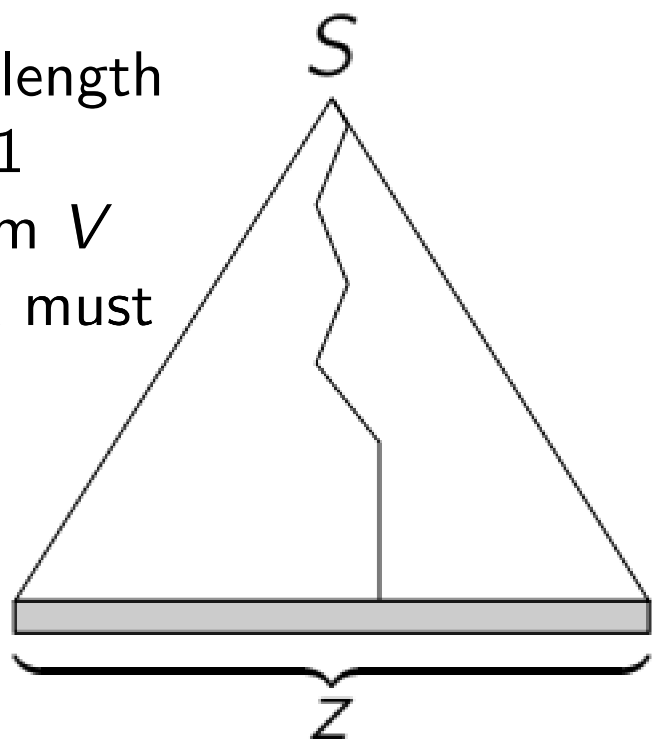


Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

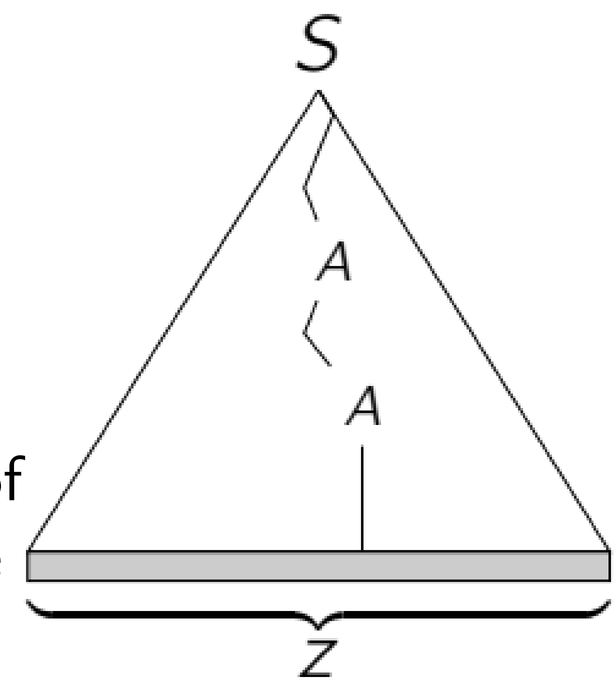
- ...
- Hence, a longest path π from root to leave has length at least $|V|$, where π consists of at least $|V| + 1$ nodes each of which is labeled by a variable from V
- As there are only $|V|$ variables, some variable A must appear twice on π



Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

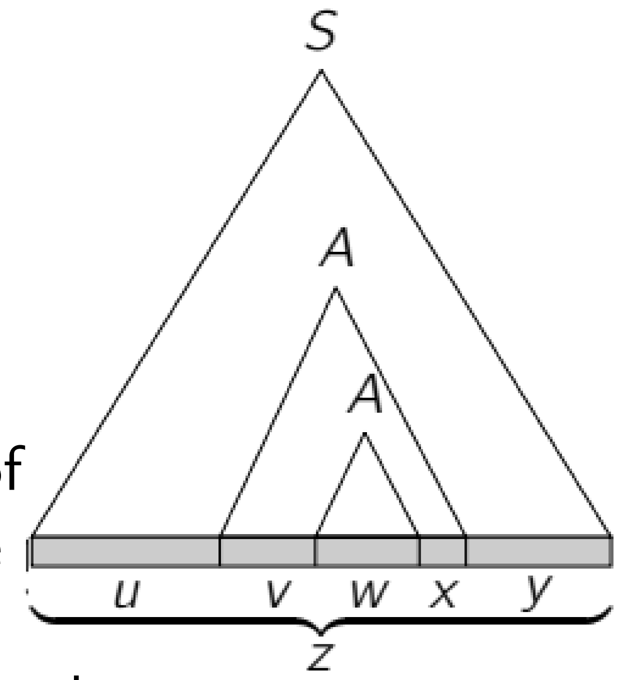


- ...
- As there are only $|V|$ variables, some variable A appears twice on π
- Consider the subtrees with root A; they allow derivation of subwords of of z.
- Subtree with lower A as root generates subword of subtree with higher A as root: $z = uvwxy$, where vwx generated by higher A and w by lower A.
- $|w| \geq 1$, as CNF grammars generate non-empty words
- $|vwx| > |w|$, as higher A above lower A, so $|vx| \geq 1$

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

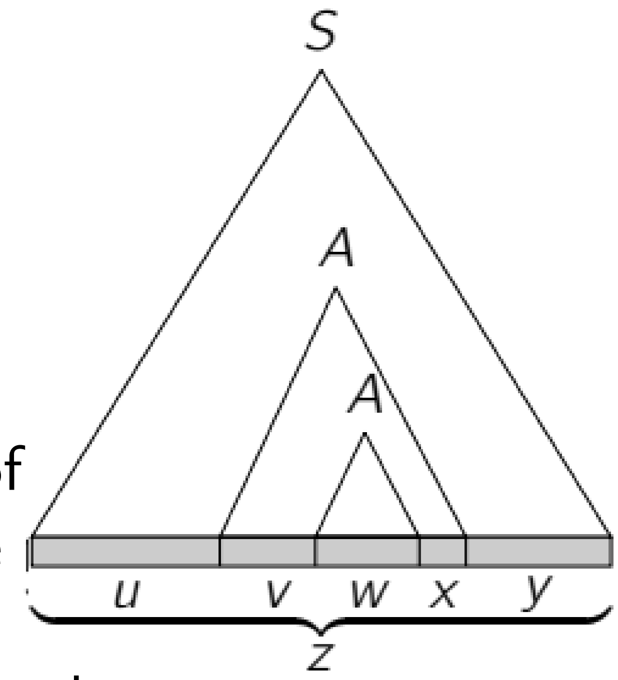


- ...
- As there are only $|V|$ variables, some variable A appears twice on π
- Consider the subtrees with root A; they allow derivation of subwords of of z.
- Subtree with lower A as root generates subword of subtree with higher A as root: $z = uvwxy$, where vwx generated by higher A and w by lower A.
- $|w| \geq 1$, as CNF grammars generate non-empty words
- $|vwx| > |w|$, as higher A above lower A, so $|vx| \geq 1$

Proof of the pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$, • $|vwx| \leq n$, • $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.



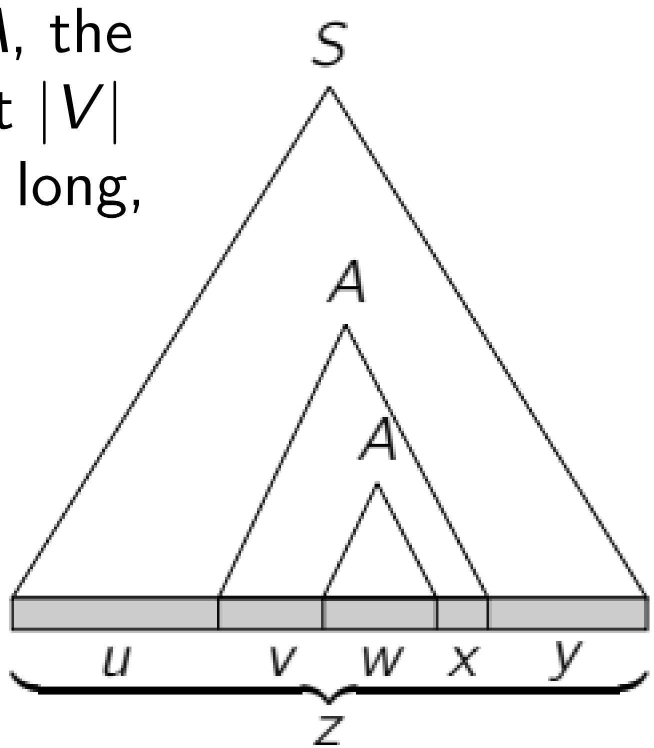
- ...
- As there are only $|V|$ variables, some variable A appears twice on π
- Consider the subtrees with root A; they allow derivation of subwords of of z.
- Subtree with lower A as root generates subword of subtree with higher A as root: $z = uvwxy$, where vwx generated by higher A and w by lower A.
- $|w| \geq 1$, as CNF grammars generate non-empty words
- $|vwx| > |w|$, as higher A above lower A, so $|vx| \geq 1$

Proof of pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$,
- $|vwx| \leq n$,
- $uv^i wx^i y \in L$ for all $i \in \mathbb{N}_0$.

- ...
- $|vwx| > |w|$, as higher A above lower A, so $|vx| \geq 1$
- As we selected the lowest occurence of higher A, the path from higher A to leaves has length at most $|V|$
- Other paths from higher A to leaves at most as long, as we chose longest path
- $\Rightarrow |vwx| \leq 2^{|V|} = n$
- Tree gives $A \rightarrow^* w$ and $A \rightarrow^* vAx$ and thus $A \rightarrow^* v^i wx^i$ for all $i \in \mathbb{N}$
- $\Rightarrow S \rightarrow^* uv^i wx^i y$ for all $i \in \mathbb{N}$

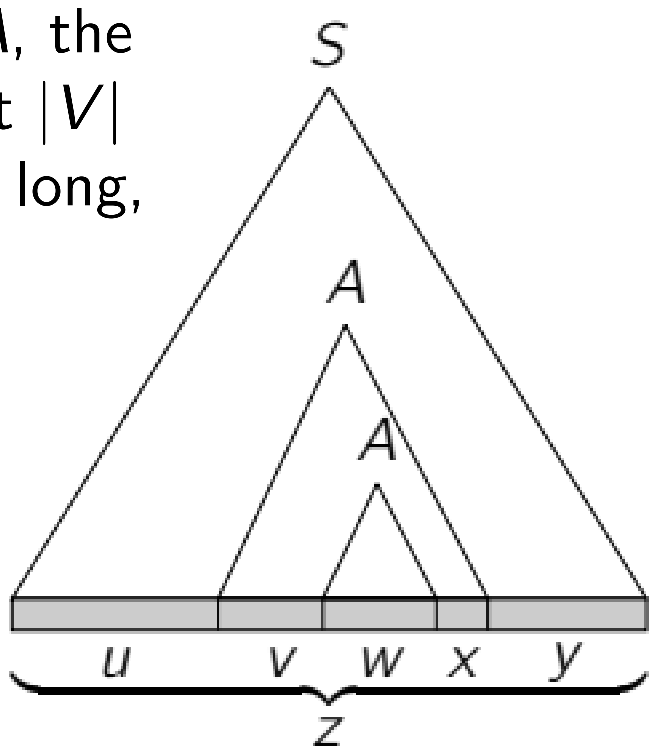


Proof of pumping lemma for CFLs

Lemma. For any context-free language L there is a number $n \in \mathbb{N}$ such that any word $z \in L$ of length $|z| \geq n$ can be decomposed as $z = uvwxy$, where

- $|vx| \geq 1$,
- $|vwx| \leq n$,
- $uv^iwx^iy \in L$ for all $i \in \mathbb{N}_0$.

- ...
- $|vwx| > |w|$, as higher A above lower A, so $|vx| \geq 1$
- As we selected the lowest occurence of higher A, the path from higher A to leaves has length at most $|V|$
- Other paths from higher A to leaves at most as long, as we chose longest path
- $\Rightarrow |vwx| \leq 2^{|V|} = n$
- Tree gives $A \rightarrow^* w$ and $A \rightarrow^* vAx$ and thus $A \rightarrow^* v^iwx^i$ for all $i \in \mathbb{N}$
- $\Rightarrow S \rightarrow^* uv^iwx^iy$ for all $i \in \mathbb{N}$



Illustrating the pumping lemma for CFLs

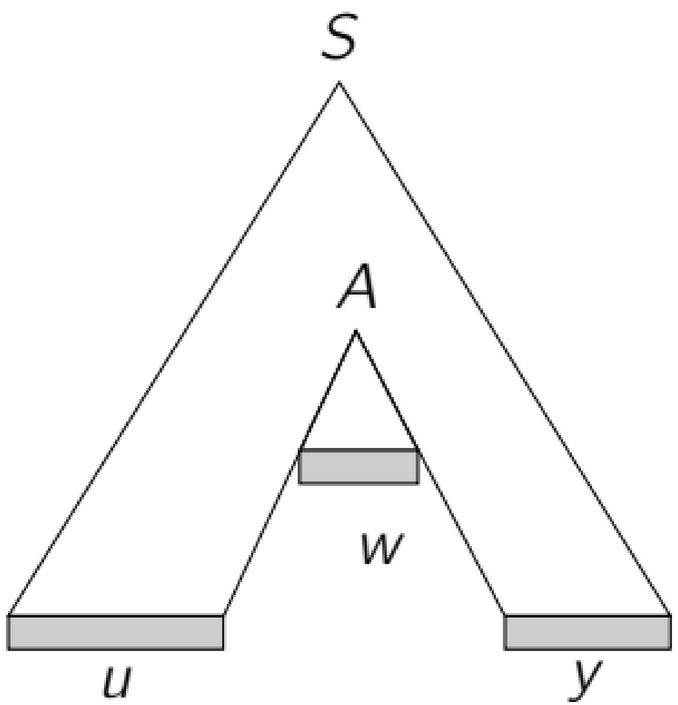


Illustration für uv^0wx^0y

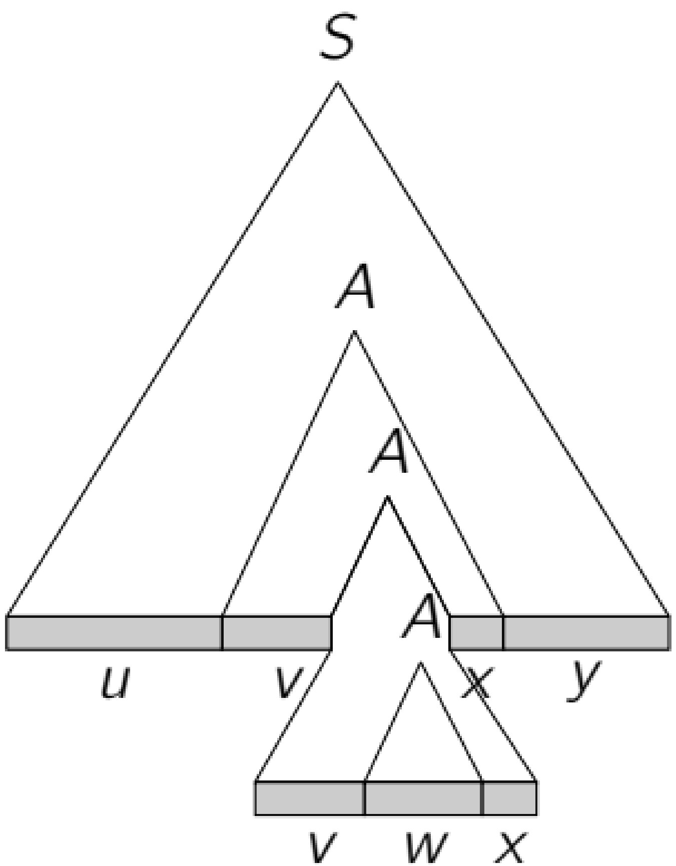


Illustration für uv^2wx^2y

Pumping lemma to refute context-freeness

- The pumping property is a necessary but not a sufficient condition of context-free languages.
- Hence the pumping lemma cannot be used (on its own) to establish the context-freeness of a language.
- However, it can be used to refute the context-freeness of a language.

Pumping lemma to refute context-freeness

Negated formulation of the pumping lemma: Consider a language L such that for any $n \in \mathbb{N}$ there is a word $z \in L$ of length at least n for which any decomposition $z = uvwxy$ with $|vx| \geq 1$ and $|vwx| \leq n$ gives an i so that $z = uv^iwx^iy \notin L$. Then L is not context-free.

Proof via translation of predicate logical formula ($\exists\forall \to \forall\exists$).

Two-player setting of pumping lemma

Let L be a language.

Two-player setting of pumping lemma

Let L be a language.

- Devil chooses number $n \in \mathbb{N}$.
- You choose word $z \in L$ with $|z| \geq n$.
- Devil chooses decomposition $z = uvwxy$ with $|wx| \geq 1$ and $|vwx| \leq n$.
- You win when $uv^iwx^iy \notin L$ for some i.

Two-player setting of pumping lemma

Let L be a language.

- Devil chooses number $n \in \mathbb{N}$.
- You choose word $z \in L$ with $|z| \geq n$.
- Devil chooses decomposition $z = uvwxy$ with $|wx| \geq 1$ and $|vwx| \leq n$.
- You win when $uv^iwx^iy \notin L$ for some i.

If you win for any choice of the devil, then L is not context-free.

Example

The language $L = \{a^i b^i c^i \mid i \in \mathbb{N}\}$ is not context-free.

Example

The language $L = \{a^i b^i c^i \mid i \in \mathbb{N}\}$ is not context-free.

- Devil selects $n \in \mathbb{N}$.
- You select $z = a^n b^n c^n$.
- Devil selects decomposition $z = uvwxy$ with $|vx| \geq 1$ and $|vwx| \leq n$.
- Case 1: vwx has form $a^i b^j$ with $i + j \leq n$. As $|vx| \geq 1$ one has $|vx|_a \geq 1$ or $|vx|_b \geq 1$, but $|vx|_c = 0$. Thus $uv^0wx^0y \notin L$.
- Case 2: vwx has form $b^i c^j$ with $i + j \leq n$. As $|vx| \geq 1$ one has $|vx|_b \geq 1$ or $|vx|_c \geq 1$, but $|vx|_a = 0$. Thus $uv^0wx^0y \notin L$.

Example application of pumping lemma

The language $L = \{a^i b^j c^i d^j \mid i, j \in \mathbb{N}\}$ is not context-free.

Example application of pumping lemma

The language $L = \{a^i b^j c^i d^j \mid i, j \in \mathbb{N}\}$ is not context-free.

- Devil chooses $n \in \mathbb{N}$.
- You choose $z = a^n b^n c^n d^n$.
- Devil chooses decomposition $z = uvwxy$ with $|vx| \geq 1$ and $|vwx| \leq n$.
- Case 1: $vwx = a^i b^j$ with $i + j \leq n$. As $|vx| \geq 1$, one has $|vx|_a + |vx|_b \geq 1$ and $uv^0wx^0y = uwy = a^{i'} b^{j'} c^n d^n$ and $i' < n$ or $j' < n$, so $uwy \notin L$.
- Case 2: $vwx = b^i c^j$ with $i + j \leq n$. As $|vx| \geq 1$, one has $|vx|_b + |vx|_c \geq 1$ and $uv^0wx^0y = uwy = a^n b^{i'} c^{j'} d^n$ and $i' < n$ or $j' < n$, so $uwy \notin L$.
- Case 3: $vwx = c^i d^j$ with $i + j \leq n$. As $|vx| \geq 1$, one has $|vx|_c + |vx|_d \geq 1$ and $uv^0wx^0y = uwy = a^n b^n c^{i'} d^{j'}$ and $i' < n$ or $j' < n$, so $uwy \notin L$.

Unary alphabet

Theorem. Let L be a language over a unary alphabet Σ, so $|\Sigma| = 1$. Then L is context-free if and only if L is regular.

Unary alphabet

Theorem. Let L be a language over a unary alphabet Σ, so $|\Sigma| = 1$. Then L is context-free if and only if L is regular.

Examples.
The languages

- $L_1 = \{a^p \mid p \text{ is prime}\}$,
- $L_2 = \{a^n \mid n \text{ is composite}\}$,
- $L_3 = \{a^n \mid n \text{ is square}\}$,
- $L_4 = \{a^{2^n} \mid n \in \mathbb{N}\}$

are not context-free.