# Learning XML : VPAs and Discrimination Trees

Cinzia Di Giusto, Davide Fissore, Etienne Lozes
Université Nice Côte d'Azur, CNRS, I3S, Sophia Antipolis, France
Stage d'été 2022

## Why VPA ?

For ∀ Non-Deterministic VPA $V_1$, there ∃ a Deterministic VPA $V_2$ such that $L(V_1) = L(V_2)$
→ Every binary operation between 2 VPA is decidable !

Note :
Push symbols ⇔ Open tags
Pop symbols ⇔ Close tags

## VPAs

VPA := Visibly pushdown automata. They can recognize context free languagages.
The alphabet is :

$$\hat{\Sigma} = \Sigma_{call} \uplus \Sigma_{ret} \uplus \Sigma_{int}$$

Acceptance for XML :
Empty stack + final states

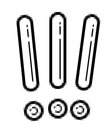
## XML

XML (eXtensible Markup Language) is a standard format for data exchange.
**XML** representable w/**VPA**!

## And Communication ?

Arthur : Does w ∈ U ?
Merlin : Yes/No

Arthur creates a conjecture C.
Arthur : Does C = U ?
Merlin : if C = U → Yes
else → a counter-example



## What is Learning ?

Dana Angluin's framework :

The Learner wants to learn a language U
The Teacher knows U

## «Canonical» VPA

Regular automata have a unique minimal (or canonical) representant, this is not true for VPA

## k-SEVPA

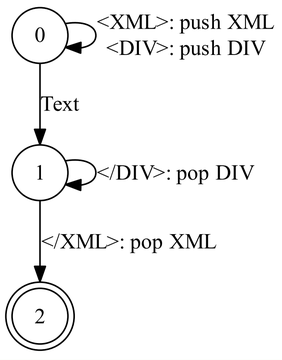
Single entry VPA are VPAs where states are partitioned into **k** modules.
Each module has only one entry for call transitions

## An XML grammar to LEARN

**G** :=
d(XML) = Text + DIV
d(DIV) = Text + DIV

d : X → <X> RULE </X>

Example:
<XML><DIV>Text</DIV></XML> ∈ **G**




## The learning phase

In Visibly Pushdown Languages (VPL), we can adapt the Myhill-Nerode congruence :
two words $(\omega_1, \omega_2) \in \hat{\Sigma}^2$ are equivalent if

$$\forall (u_1, u_2) \in \mathrm{WM}(\hat{\Sigma})$$
$$u_1 \cdot \omega_1 \cdot u_2 \in L \leftrightarrow u_1 \cdot \omega_2 \cdot u_2 \in L$$

$\mathrm{WM}(\hat{\Sigma})$ : It is a couple of words, called well-matched words, $u_1$, $u_2$ such that every call symbol of $u = u_1 \circ u_2$ has a corresponding ret symbol

## Discrimination Tree

Thanks to Well-Matched words, we can build the Discrimination tree :

- Inner Nodes contain a couple ($u_1$, $u_2$) forming a WM
- Leaves are labelled with a string.

## Leaves meaning

Leaves represent the states of the VPA and are determinded through Membership queries

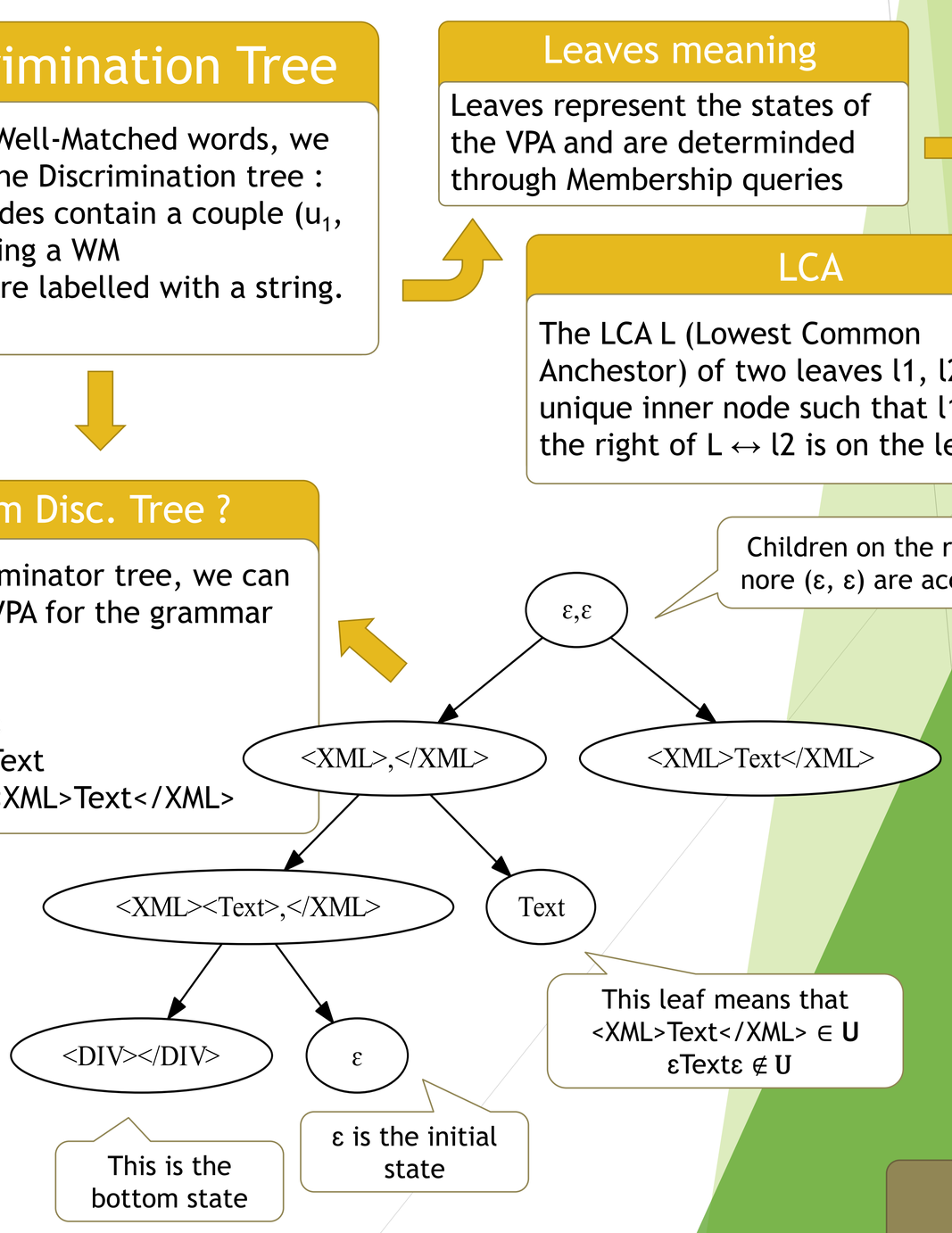
## LCA

The LCA L (Lowest Common Anchestor) of two leaves l1, l2 is the unique inner node such that l1 is on the right of L ↔ l2 is on the left of L

## VPA from Disc. Tree ?

From this discriminator tree, we can buildthe same VPA for the grammar **G**. Where:

state 0 := leaf ε
state 1 := leaf Text
state 2 := leaf <XML>Text</XML>




## Demo


## References

.bib

*This work has been supported by the French government, through the UCA DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-17-EURE-0004."*