Model-Checking Parse Trees

A. Boral S. Schmitz

CMI, Chennai LSV, ENS Cachan

LICS 2013, June 26th 2013

Modeling Syntax (with Trees)

(e.g. Pullum and Scholz, 2001)

generative syntax context-free grammars, ...

model-theoretic syntax MSO, $\mathrm{PDL}_{\mathrm{tree}}$, ...

mixed here for grammar engineering

Modeling Syntax (with Trees)

(e.g. Pullum and Scholz, 2001)

generative syntax context-free grammars, ...

model-theoretic syntax MSO, $\mathrm{PDL}_{\mathrm{tree}}$, ...

mixed here for grammar engineering

Outline

Recognition Across Formalisms

CFG	local specification	PTime-c.
$\mathrm{PDL}_{\mathrm{tree}}$	long-distance navigation	ExpTime-c.
MSO	global specification	Tower-c.

Parse Forest Model-Checking

input a context free grammar G, a word w, and a $\mathrm{PDL}_{\mathsf{tree}}$ formula φ,

question does there exists a tree $t \in L_{G,w}$ s.t. $t \models \varphi$?

PDL on Trees

(Afanasiev et al., 2005)

- finite, ordered, AP-labeled trees
- two relations: child '$\downarrow$', right-sibling '$\rightarrow$'
- syntax of $\mathrm{PDL}_{\mathrm{tree}}$ (aka Regular XPath):

$$\varphi ::= p \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\pi\rangle\varphi \quad \text{(node formulæ)}$$

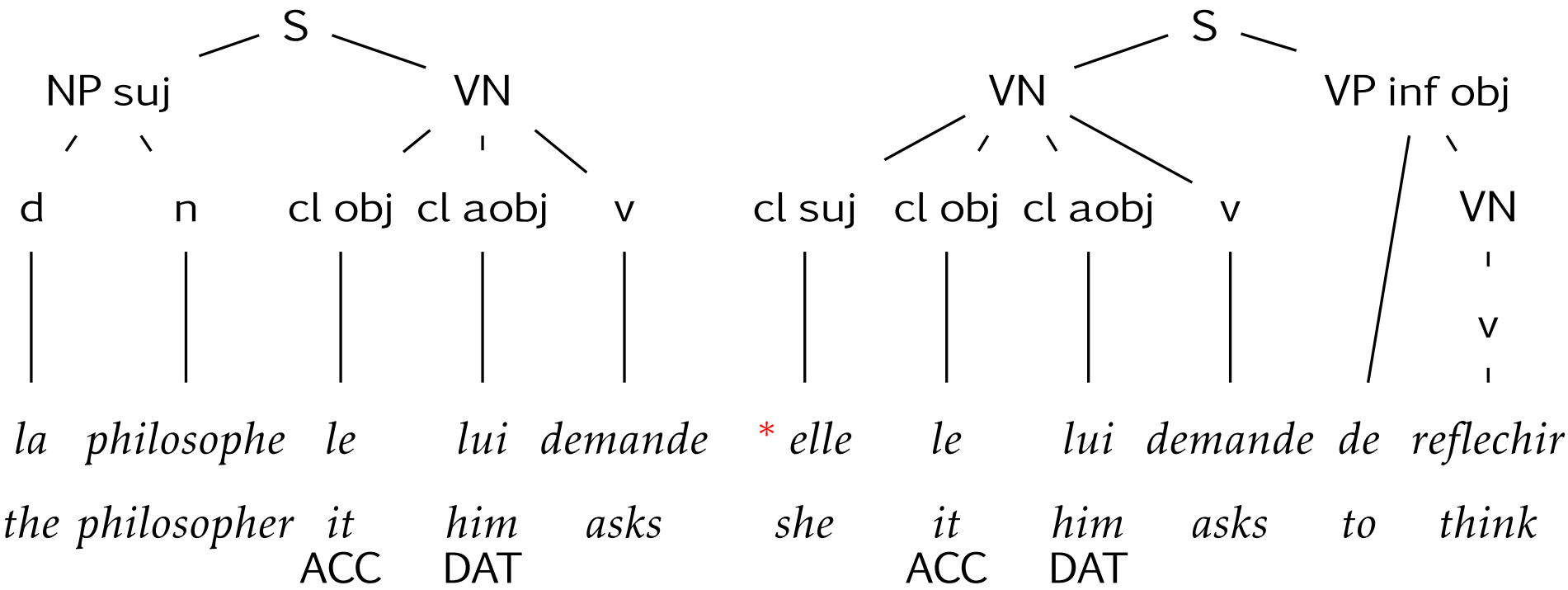
$$\pi ::= \downarrow \mid \rightarrow \mid \pi\,;\pi \mid \pi + \pi \mid \pi^* \mid \pi^{-1} \mid \varphi? \quad \text{(path formulæ)}$$

Long-Distance Dependencies

- local syntax using a CFG
- long-distance constraints enforced through a $\mathrm{PDL}_{\mathrm{tree}}$ formula

Example: French Clitics

$$S \rightarrow NPsuj?\ VN\ VPinfobj?\ PPaobj?$$
$$VN \rightarrow clsuj?\ clobj?\ claobj?\ v$$



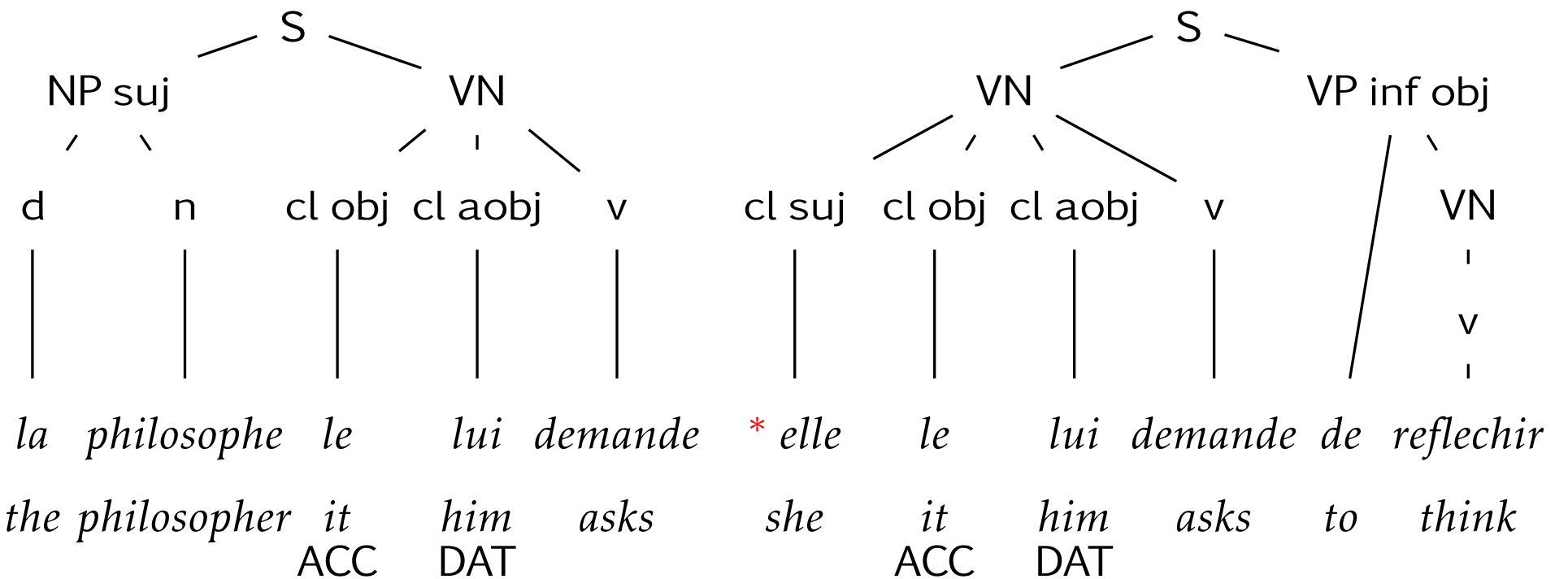
Example: French Clitics

$$[\downarrow^*]demande \Rightarrow \big(\langle(\uparrow;\uparrow;\rightarrow^+) + (\uparrow;\leftarrow^+;\mathrm{cl?})\rangle \mathrm{obj} \quad \text{(at least one object)}$$

$$\wedge \langle(\uparrow;\uparrow;\leftarrow) + (\uparrow;\leftarrow^+;\mathrm{cl?})\rangle \mathrm{suj} \quad \text{(at least one subject)}$$

$$\wedge \bigwedge_{f\in\{\mathrm{suj,obj,aobj}\}} \langle\uparrow;\leftarrow^+;\mathrm{cl?}\rangle f \Rightarrow \neg\langle\uparrow;\uparrow;(\leftarrow + \rightarrow^+)\rangle f\big)$$

(clitic arguments forbid the matching canonical arguments)



Discussion

- grammar
  - ε-free: $A \to \varepsilon$ forbidden
    - acyclic: $A \Rightarrow^{+} A$ forbidden
- formula
  - here $\mathrm{PDL}_{\mathsf{core}}$ (aka Core XPath)
    - in the literature: $\mathrm{PDL}_{\mathsf{cp}}$ (aka Conditional XPath, see Palm, 1999; Lai and Bird, 2010)
- model
  - richer structures than trees?
    - trace can be emulated by a formula

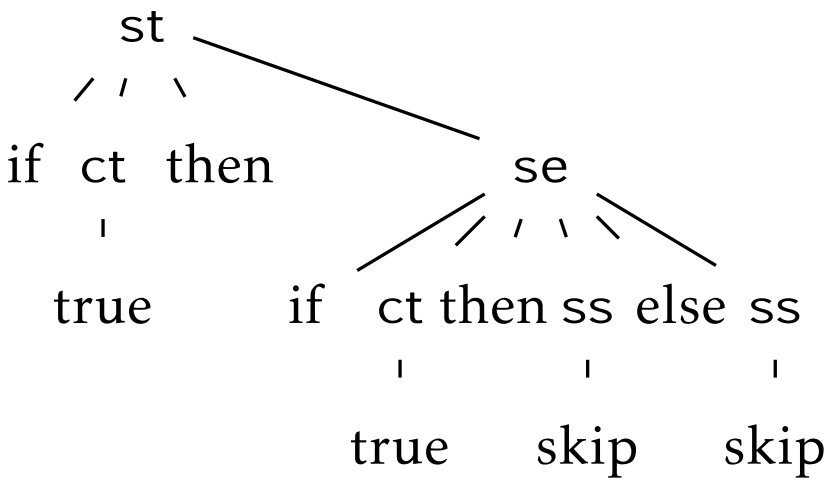
Ambiguity Filtering

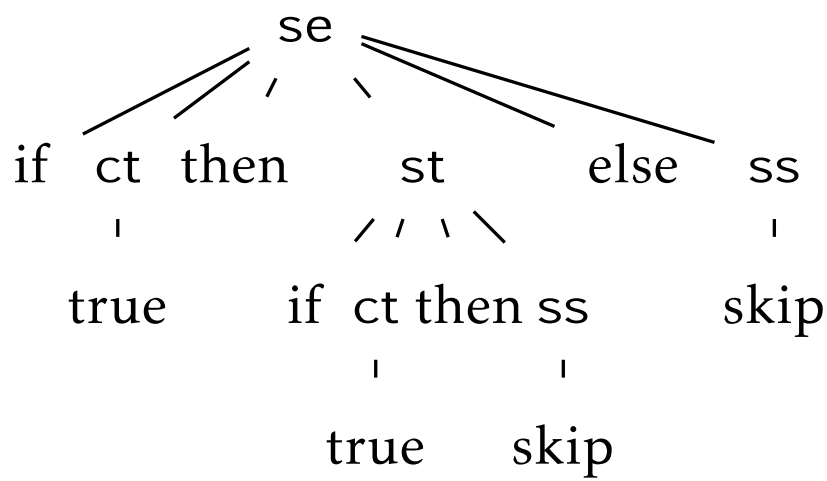
(e.g. Klint and Visser, 1994)

- syntax defined through a CFG
- exclude spurious parse trees, here thanks to a formula

Example: Dangling Else

$$S \to \mathrm{st}(\text{if } C \text{ then } S) \mid \mathrm{se}(\text{if } C \text{ then } S \text{ else } S)$$
$$\mid \mathrm{sw}(\text{while } C\ S) \mid \mathrm{ss}(\text{skip})$$
$$C \to \mathrm{ct}(\mathrm{true}) \mid \mathrm{cf}(\mathrm{false})$$





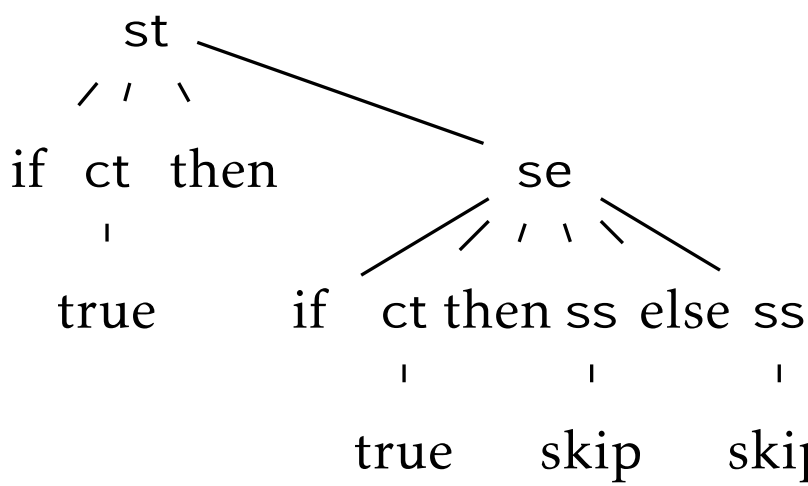
Example: Dangling Else

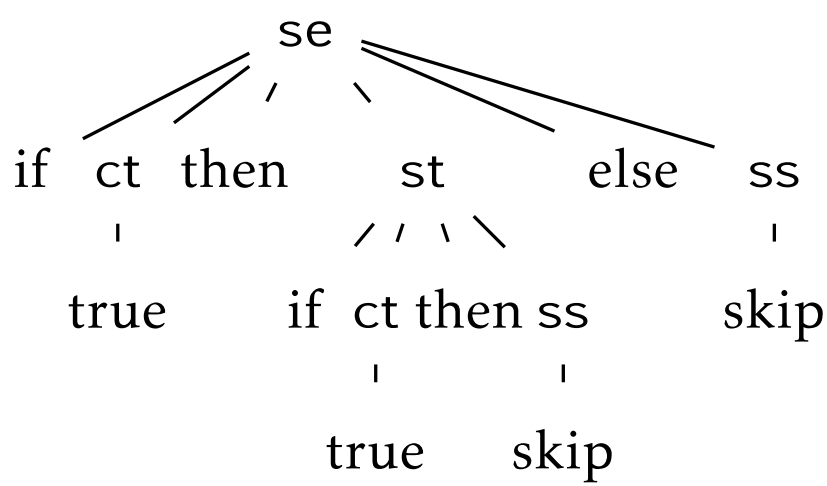
$$\mathsf{first} \stackrel{\text{def}}{=} \neg\langle\leftarrow\rangle\top$$

$$\mathsf{last} \stackrel{\text{def}}{=} \neg\langle\rightarrow\rangle\top$$

$$\prec \stackrel{\text{def}}{=} (\mathsf{last}?;\uparrow)^*;\rightarrow;(\downarrow;\mathsf{first}?)^*$$

$$\varphi = \neg\langle\downarrow^*\rangle(\mathsf{st} \wedge \langle\prec\rangle\mathsf{else})$$





Discussion

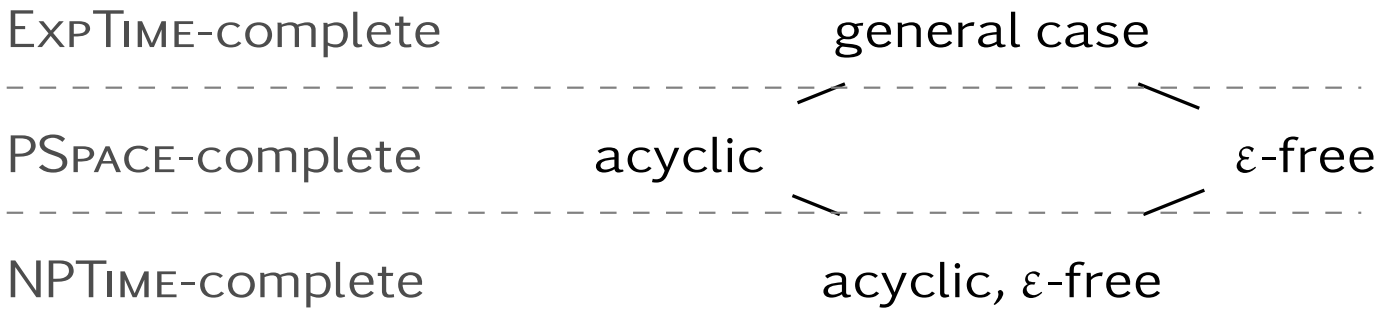
grammar ▸ acyclic grammars

formula ▸ here in $\mathrm{PDL}_{\mathsf{cp}}$ (but also in XPath 1.0)

model ▸ "static" filtering: can be compiled into the grammar

- handling layout-sensitive syntax (aka offside rules)? (c.f. Adams, 2013)

Complexity Summary



- borrow from similar problems in XML theory by Benedikt et al. (2008)
- lower bounds already hold for the $\mathrm{PDL}_{\mathsf{core}}[\downarrow]$ fragment, often for fixed G and/or w.

Parse Forest $L_{G,w}$

Theorem (Bar-Hillel et al., 1961)

Let G be a CFG and w a string. Then we can construct in polynomial time a tree automaton $\mathcal{A}_{G,w}$ of polynomial size, s.t. $L(\mathcal{A}_{G,w}) = L_{G,w}$.

- intuition: the states of $\mathcal{A}_{G,w}$ are triples (i, A, j) where $0 \leqslant i \leqslant j \leqslant |w|$ and $w_i \cdots w_j \in L_G(A)$.

- cardinality of $L_{G,w}$: potentially

Parse Forest $L_{G,w}$

Theorem (Bar-Hillel et al., 1961)

Let G be a CFG and w a string. Then we can construct in polynomial time a tree automaton $\mathcal{A}_{G,w}$ of polynomial size, s.t. $L(\mathcal{A}_{G,w}) = L_{G,w}$.

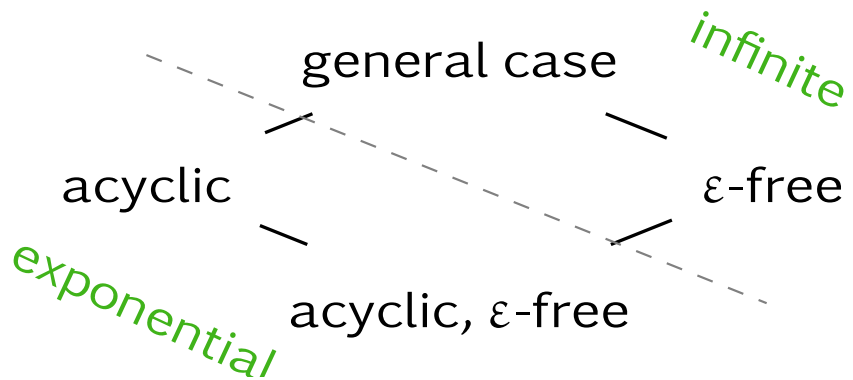
- intuition: the states of $\mathcal{A}_{G,w}$ are triples (i, A, j) where $0 \leqslant i \leqslant j \leqslant |w|$ and $w_i \cdots w_j \in L_G(A)$.
- cardinality of $L_{G,w}$: potentially

Parse Forest $L_{G,w}$

Theorem (Bar-Hillel et al., 1961)

Let G be a CFG and w a string. Then we can construct in polynomial time a tree automaton $\mathcal{A}_{G,w}$ of polynomial size, s.t. $L(\mathcal{A}_{G,w}) = L_{G,w}$.

- intuition: the states of $\mathcal{A}_{G,w}$ are triples (i, A, j) where $0 \leqslant i \leqslant j \leqslant |w|$ and $w_i \cdots w_j \in L_G(A)$.
- cardinality of $L_{G,w}$: potentially



Zoom on PSpace Results

- build a nonrecursive DTD D from G and w
  - DTD: "extended" CFG with regular right parts
  - thus unranked trees
  - nonrecursive: $A \Rightarrow^+ \alpha A \beta$ forbidden
- build a formula φ'

Theorem (Benedikt et al., 2008)
Satisfiability of a $\mathrm{PDL}_{\mathsf{core}}$ *formula* φ' *in presence of a nonrecursive DTD* D *is in* PSpace.

(Reduces to emptiness of loop-free alternating two-ways word automata on an XML encoding (Serre, 2006).)

Zoom on PSpace Results

- build a nonrecursive DTD D from G and w
  - DTD: "extended" CFG with regular right parts
  - thus unranked trees
  - nonrecursive: $A \Rightarrow^+ \alpha A \beta$ forbidden
- build a formula φ'

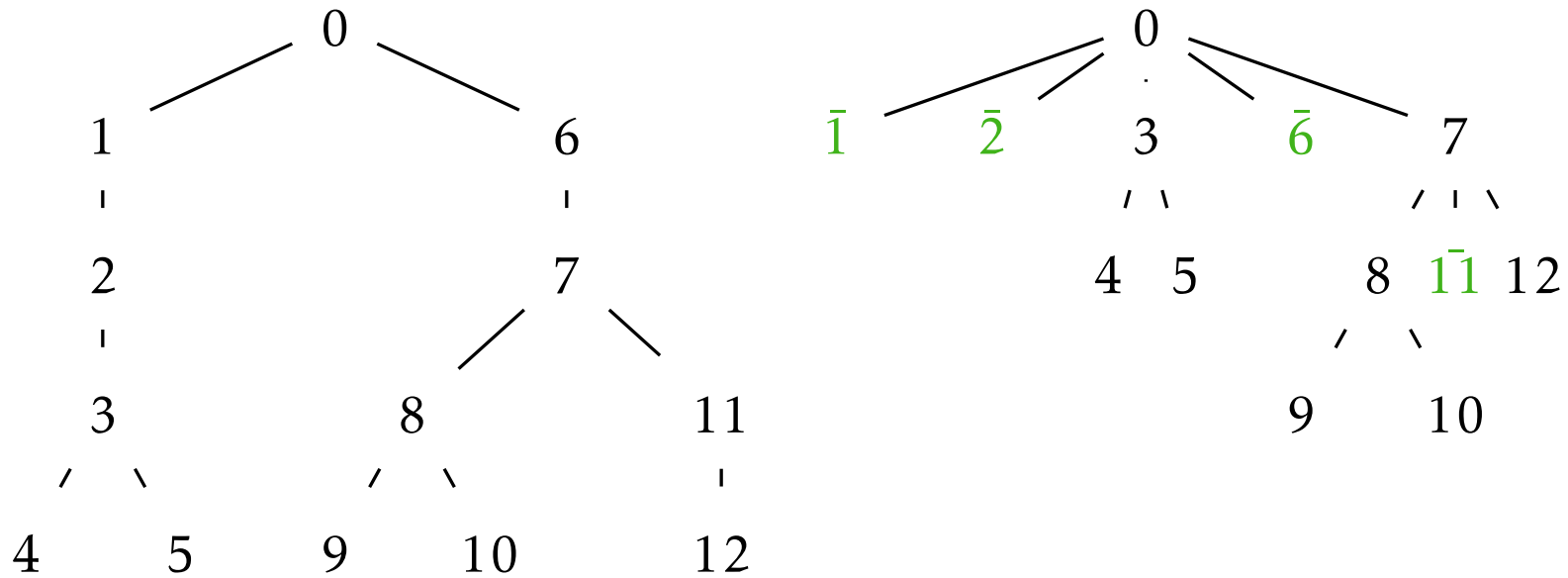
Theorem

Satisfiability of a $\mathrm{PDL}_{\mathsf{tree}}$ *formula* φ' *in presence of a nonrecursive DTD* D *is in* PSpace.

(Reduces to emptiness of alternating parity two-ways word automata on an XML encoding (Serre, 2006).)

Reduction in the ε-free Case

- $L_{G,w}$ can be infinite
- chains of transitions in $\mathcal{A}_{G,w}$: $q_1 \Rightarrow q_2 \Rightarrow \cdots \Rightarrow q_n$
- maximal chain if q_n binary or a leaf
- "rotate" chains



Reduction in the ε-free Case (cont'd)

- chains(q) the set of maximal chains starting from q forms a regular substitution
- D obtained from $\mathcal{A}_{G,w}$ by applying chains(q)
- φ has to be interpreted on the "rotated" trees

Proposition

PFMC with ε-free grammars is in PSpace.

Corollary

$\mathrm{PDL}_{\mathrm{tree}}$ *recognition with ε-free models is in* PSpace.

Concluding Remarks

- mixed approach to syntax:
  - bulk work with a CFG
  - fine tuning with a $\mathrm{PDL}_{\mathrm{tree}}$(or MSO, or...) formula
- new applications for model-checking techniques
  - different practical restrictions
  - leading to improved complexities
- exploit data logics?

References

Adams, M.D., 2013. Principled parsing for indentation-sensitive languages: revisiting Landin's offside rule. In *POPL 2013*, pages 511–522. doi:10.1145/2429069.2429129.

Afanasiev, L., Blackburn, P., Dimitriou, I., Gaiffe, B., Goris, E., Marx, M., and de Rijke, M., 2005. PDL for ordered trees. *J. Appl. Non-Classical Log.*, 15(2):115–135. doi:10.3166/jancl.15.115-135.

Bar-Hillel, Y., Perles, M., and Shamir, E., 1961. On formal properties of simple phrase-structure grammars. *Z. Phonetik Sprachwiss. Kommunik.*, 14:143–172.

Benedikt, M., Fan, W., and Geerts, F., 2008. XPath satisfiability in the presence of DTDs. *Journal of the ACM*, 55(2:8). doi:10.1145/1346330.1346333.

Klint, P. and Visser, E., 1994. Using filters for the disambiguation of context-free grammars. In Pighizzini, G. and San Pietro, P., editors, *ASMICS Workshop on Parsing Theory*, Technical Report 126-1994, pages 89–100. Università di Milano.

Lai, C. and Bird, S., 2010. Querying linguistic trees. *J. Logic Lang. Inform.*, 19(1):53–73. doi:10.1007/s10849-009-9086-9.

Palm, A., 1999. Propositional tense logic of finite trees. In *MOL 6*.

Pullum, G.K. and Scholz, B.C., 2001. On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *LACL 2001*, volume 2099 of *LNCS*, pages 17–43. Springer. doi:10.1007/3-540-48199-0_2.

Serre, O., 2006. Parity games played on transition graphs of one-counter processes. In Aceto, L. and Ingólfsdóttir, A., editors, *FoSSaCS 2006*, volume 3921 of *LNCS*, pages 337–351. Springer. doi:10.1007/11690634_23.