Around dot depth two

Kamal Lodaya

The Institute of Mathematical Sciences, Chennai

Joint work with Paritosh Pandya and Simoni Shah (TIFR, Mumbai)

Regular languages

Language L over Σ^*.

- Recognized by automata
- Defined by regular expressions
- Specified by a formula of classical or temporal logic
  - Every a is eventually followed by b
  - $\mathsf{G}(a \Rightarrow \mathsf{F}\, b)$
  - $\forall x.\ (a(x) \Rightarrow \exists y.\ (x < y \wedge b(y)))$
- Specified algebraically by its syntactic monoid.

Expressiveness Language classes and their corresponding expressions/logics/algebra/automata.

Subclasses of regular languages

Some well known results.

Logic	Interval TL	TL	Automata	Algebra
$MSO[<]$	QDDC	QPTL	DFA	Monoids
$FO[<]$, $FO^3[<]$	ITL	LTL[X,Y,U,S]	CFA	Aperiodic Monoids
$FO^2[<, S]$	?	LTL[X,Y,F,P]	?	DA*D
$FO^2[<]$	?	LTL[F,P]	$po2dfa$	DA

Answer

- A fragment of ITL/SF with left- and right-deterministic marked concatenation which we call Unambiguous interval temporal logic/ Unambiguous starfree expressions
- An extension of $po2dfa$ which look around which we call $po2dla$

More questions

We have worked out our results in the framework of interval temporal logic and believe that a similar story can be told for starfree expressions.

- Membership for SF/ITL is Ptime-complete (Petersen)
- Satisfiability for QDDC and even for ITL is nonelementary (Stockmeyer and Meyer)
- What about membership and satisfiability for unambiguous interval temporal logic?

Motivation Modelchecker DCVALID worked more efficiently because formulas enhanced with "marking" propositions generated small automata (Krishna and Pandya)

Example Let $w = \triangleright ddabdabddcb \triangleleft$. Matched by the pattern $b\underline{d}d$.

Logic ITL

- Word $w \in \Sigma^+$
- $pos(w)$ set of positions in the word.
- $INTV(w) \stackrel{\text{def}}{=} \{[i,j] \in pos(w)^2 \mid i \leq j\}$
- Satisfaction $w, [i,j] \models D$
- $w \models D$ iff $w, [1, \#w] \models D$.
  $L(D) = \{w \in \Sigma^+ \mid w \models D\}$

Syntax and semantics

Let $a \in \Sigma$. Let D, D_1, D_2 range over formulas in ITL.

Abstract syntax of ITL:

$$pt \;\mid\; \lceil a \rceil \;\mid\; D_1 ; D_2 \;\mid\; D_1 \vee D_2 \;\mid\; \neg D$$

Semantics of ITL:

$$w,[i,j] \models pt \text{ iff } i = j$$

$$w,[i,j] \models \lceil a \rceil \text{ iff for all } k : i \leq k \leq j : w[i] = a$$

$$w,[i,j] \models D_1 ; D_2 \text{ iff for some } k : i \leq k \leq j \text{ and}$$
$$w,[i,k] \models D_1 \text{ and } w,[k,j] \models D_2$$

Results

TL	Satisfiability	Interval TL	Membership	Satisfiability
LTL[X,U]	Pspace	ITL	Pspace	Nonelementary
LTL[X,Y,F,P]	Pspace	LITL	LogDCFL	NP/Pspace
LTL[F,P]	NP	UITL	LogDCFL	NP

- ITL subclasses expressively equivalent to FO^2
- Direct effective translations from interval logics to $po2dla$, $po2dfa$(unlike FO^2, unary LTL)

UITL results presented at IFIP TCS Conference, Milano, Sep '08

Unambiguous languages

- Studied by Schützenberger (1976)
- A concatenation e_1e_2 is *unambiguous* if for any word w in $L(e_1e_2)$, there is a unique factorization $w = uv$ such that $u \in L(e_1)$ and $v \in L(e_2)$
- An unambiguous language is recognized by a finite monoid which satisfies for some n the equation $(xyz)^{2n} = (xyz)^n y (xyz)^n$
- Such monoids are said to belong to the pseudovariety DA
- Conversely, every monoid in DA recognizes an unambiguous language
- Unambiguous languages are precisely those definable in $FO^2[<]$ (Thérien and Wilke 1998)

Po2DFA which look around

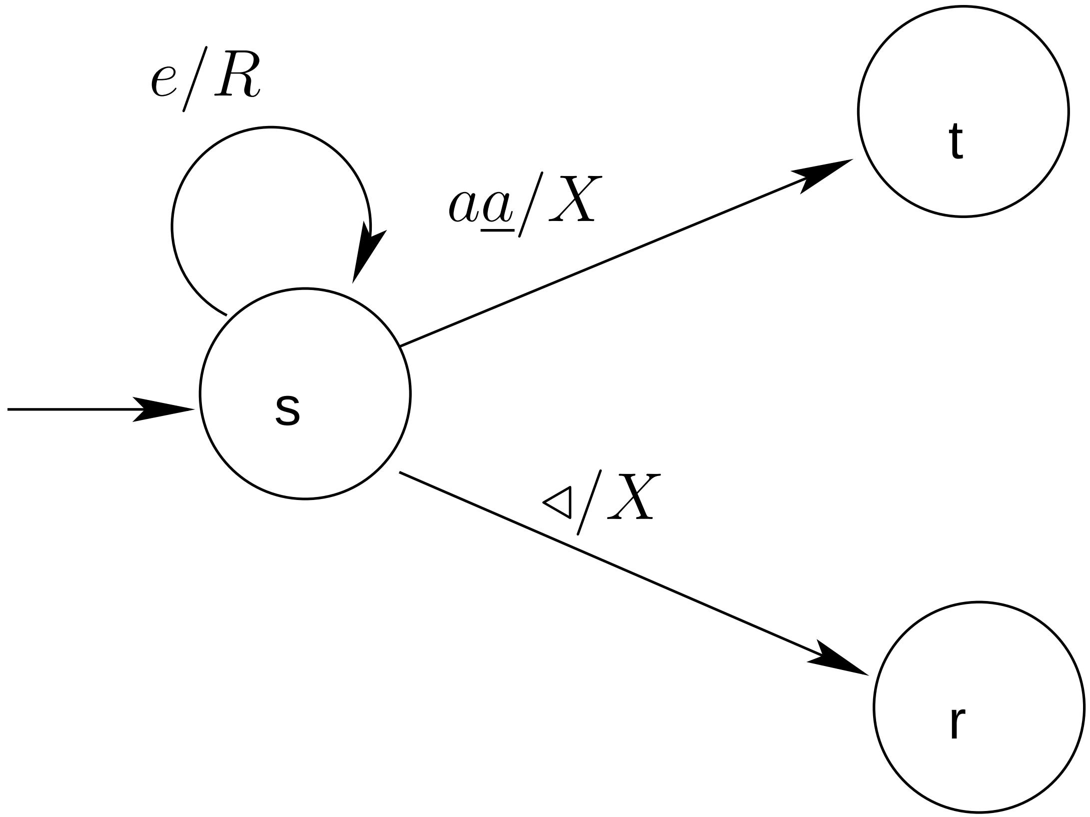
Generalize partially ordered two-way DFA introduced by Schwentick, Thérien and Vollmer (2001).

A *po2dla* $M = (Q, \leq, \delta, s, t, r)$ over $\Sigma \cup \{\triangleright, \triangleleft\}$ where

- $(Q, \leq)$ is a poset of states with s (start), r (reject) and t (accept), the latter two minimal elements
- $\delta(q, u\underline{a}v) = (q', \_) \in Q \times \{L, R, X\}$ implies $q' < q$ (once a state is exited, it is not re-entered)
- $\delta(q, e) = (q', \_) \in Q \times \{L, R\}$ implies $q' \leq q$ (else skip to the next letter)
- Endmarkers to prevent falling off the end
- Overlapping of $u\underline{a}v$'s disallowed for determinism
- Maximum lookaround ($|u|$ or $|v|$) is k

Example $po2dla$

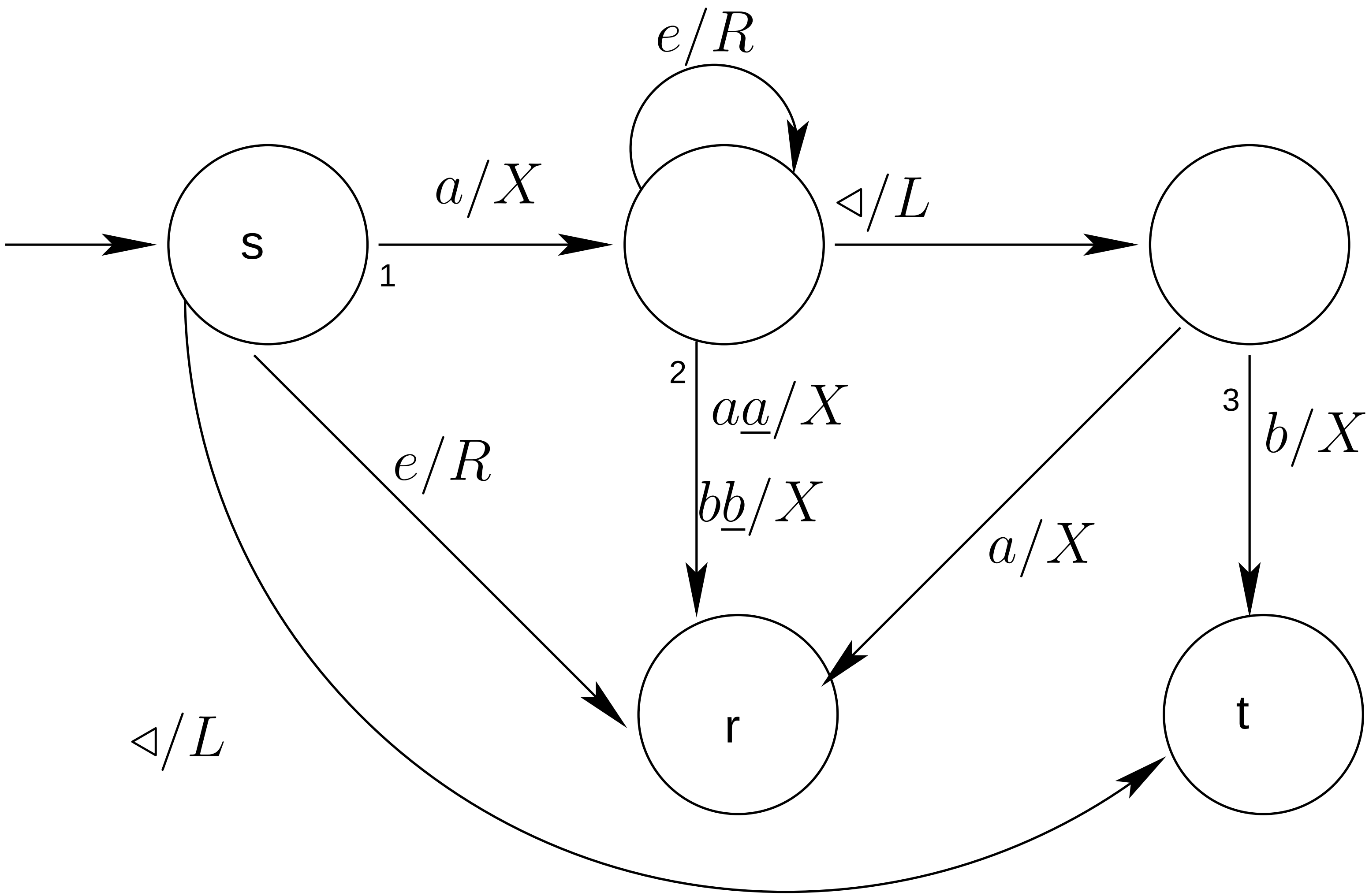
Accepting the language $\Sigma^* aa \Sigma^*$.



Example $po2dla$

Accepting the language $(ab)^*$.

Transition 1 checks if the word begins with a, transition 2 rejects the word if there are consecutive $a's$ or $b's$, and transition 3 accepts if it ends with a b.



Results for $po2dla$

Consider a $po2dla$ with n states, lookaround k and a word of length m.

- Membership in time $nm + 2k + 1$, and in space linear in k and logarithmic in m, n
- Small model of size $n|\Sigma|^{2k+1}$
- Nonemptiness in Pspace (NP for fixed lookaround)
- Boolean operations cause polynomial blowup in size: compositional way of constructing $po2dla$, extending the turtle expressions originally defined by Schwentick, Thérien and Vollmer (2001).

Extended turtle expressions

- Acc accepts and Rej rejects without moving the head.
- Scan for a pattern $u\underline{a}v$ while *disallowing* another pattern (or a specified set of patterns)
- $P?Q, R$ executes P first. If P accepts at position j then Q is executed from head position j. Otherwise R is executed from head position j.

Logic $LITL$

Let $a \in \Sigma$. Let D, D_1, D_2 range over formulas in $LITL$.

Abstract syntax of $LITL$:

$$\top \mid pt \mid D_1 F_{u\underline{a}v} D_2 \mid D_1 L_{u\underline{a}v} D_2$$
$$\mid D_1 \vee D_2 \mid \neg D \mid \oplus D \mid \ominus D$$

Examples $\top F_{\underline{a}a} \top$ holds if there is an occurrence of the factor aa

$\neg(\top L_{b\underline{b}} \top) F_{\underline{a}a} \top$ says that the factor bb does not occur before the first occurrence of the factor aa

Semantics

$w, [i, j] \models D_1 F_{u\underline{a}v} D_2$ iff for some $k : i \leq k \leq j.\ w[*, k, *] = uav$
and (for all $m : i \leq m < k.\ w[*, m, *] \neq uav$)
and $w, [i, k] \models D_1$ and $w, [k, j] \models D_2$

$w, [i, j] \models D_1 L_{u\underline{a}v} D_2$ iff for some $k : i \leq k \leq j.\ w[*, k, *] = uav$
and (for all $m : k < m \leq j.\ w[*, m, *] \neq uav$)
and $w, [i, k] \models D_1$ and $w, [k, j] \models D_2$

$w, [i, j] \models \oplus D$ iff $i < j$ and $w, [i + 1, j] \models D$

$w, [i, j] \models \ominus D$ iff $i < j$ and $w, [i, j - 1] \models D$

Results for $LITL$

- Automaton characterization: $po2dla$ are partially ordered two-way DFA which can check factors in their context (extending ideas used for locally testable languages).
- Membership in LogDCFL, emptiness in Pspace (NP for fixed lookaround)
- Ptime construction $A: \ LITL \rightarrow po2dla$ such that $L(D) = L(A(D))$. The size $|A(D)| \ = \ \mathcal{O}(|D|^2)$ states and same lookaround size as formula.
- Exptime construction $D: \ po2dla \rightarrow LITL$ such that $L(A) = L(D(A))$. The size $|D(A)| \ = \ \mathcal{O}(2^{|A|})$ and same lookaround size as automaton.

Unambiguous interval temporal logic

Obtained by disallowing lookahead, a instead of $u\underline{a}v$

Property Between the last a and subsequent first d there must be no b.

Language Let $\Sigma = \{a, b, c, d\}$

$$\Sigma^* a c^* d \{b, c, d\}^*$$

Formula $D \stackrel{\text{def}}{=} (\top L_a((\neg(\top F_b \top))F_d \top))$

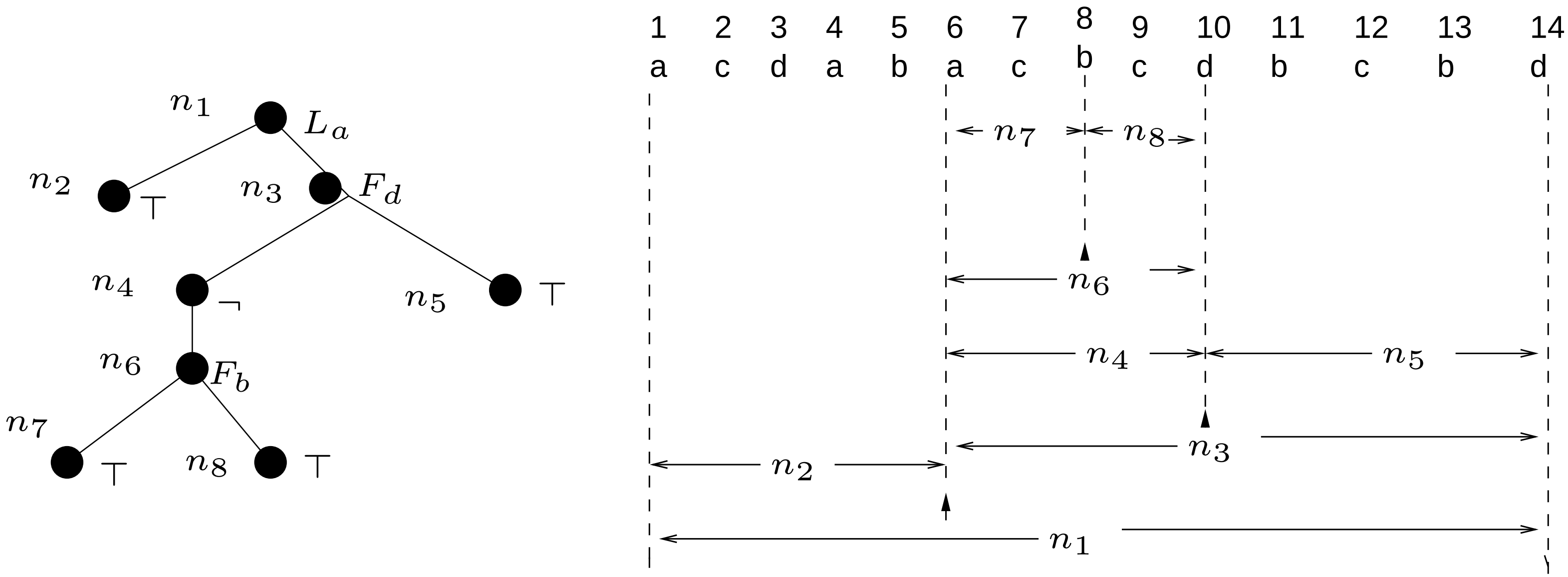
Let $a \in \Sigma$. Let D, D_1, D_2 range over formulas in $UITL$. (We use $\top$ for "true").

Abstract syntax of $UITL$:

$$\top \mid pt \mid D_1 F_a D_2 \mid D_1 L_a D_2 \mid D_1 \vee D_2 \mid \neg D \mid \oplus D \mid \ominus D$$

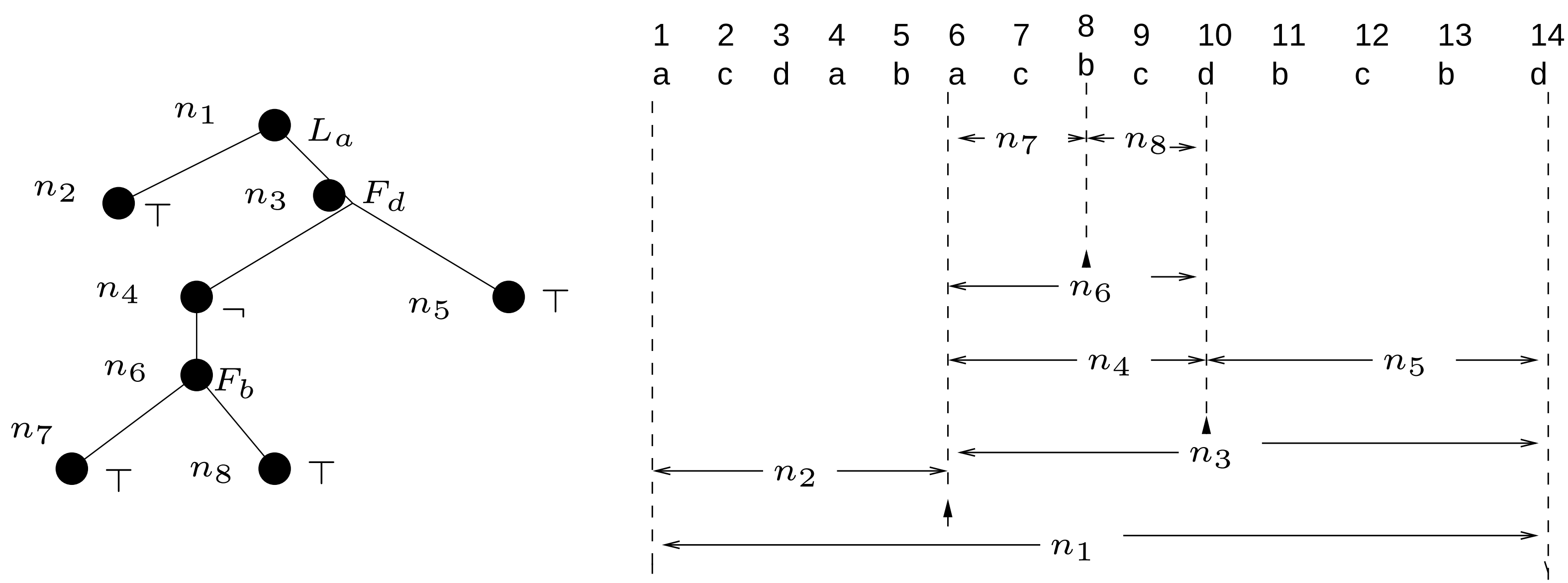
Example

Formula $D \stackrel{\text{def}}{=} (\top L_a((\neg(\top F_b \top))F_d \top))$



Example

Formula $D \stackrel{\text{def}}{=} (\top L_a ((\neg(\top F_b \top)) F_d \top))$



$w, [6, 14] \models n_4\ F_d\ n_5$ iff $w, [6, 10] \models n_4$ and $w, [10, 14] \models n_5$.

Unique parsability

Given a word w,

- Each node has unique associated interval (or none).

  $$Intv_w : Nodes \rightarrow INTV(w) \cup \{\mathbf{u}\}$$

  $Intv_w(n)$ depends only on the context of n.

- Each node of the form F_a or L_a has a unique “chopping” position (or none).

  $$cPos_w : MNodes \rightarrow INTV(w) \cup \{\mathbf{u}\}$$

- Truth of a node in its unique interval

  $$Eval_w : Nodes \rightarrow \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}.$$

Results for $UITL$

- Membership in LogDCFL, satisfiability in NP
- Ptime construction $A : \ UITL \rightarrow po2dfa$ such that $L(D) = L(A(D))$. The size $|A(D)| \ = \ \mathcal{O}(|D|^2)$ states
- Exptime construction $D : \ po2dfa \rightarrow UITL$ such that $L(A) = L(A(D))$. The size $|D(A)| \ = \ \mathcal{O}(2^{|A|})$.

Subformula automaton

Theorem For each node n we construct $po2dfa$ $\mathcal{M}(n)$ such that

- If $Eval_w(n) = t$ then $\mathcal{M}(n)(w, *) = (t, j)$ for some j.
- If $Eval_w(n) = f$ then $\mathcal{M}(n)(w, *) = (r, j)$ for some j.

Construction of $\mathcal{M}(n)$ By structural Induction.

- $\mathcal{M}(\top) \quad = \quad Acc$.
- $\mathcal{M}(\neg n) \quad = \quad \mathcal{M}(n)?Rej, Acc$.
- $\mathcal{M}(n_1 \vee n_2) \quad = \quad \mathcal{M}(n_1)?Acc, \mathcal{M}(n_2)$.

Automaton for F_a construct

Let $n = n_1 \; F_a \; n_2$. Then $\mathcal{M}(n) \;\; =$

Reach left boundary of n; Forward scan for a;
Check if head is in $Intv_w(n)$;
$\mathcal{M}(n_1); \;\; \mathcal{M}(n_2)$

Main difficulty: How to check if current head position is in $Intv_w(n)$?
Proof idea: Carry the context around (the sequence of F_a and L_a operators going up to the root from node n) using a stack while doing the automaton construction.

Context lemma: Matching the context in the stack can be done by a linear-sized extended turtle expression.

Converse

- Effective construction $D: \ po2dfa \rightarrow UITL$ such that $L(A) = L(D(A))$. The size $|D(A)| \ = \ \mathcal{O}(2^{|A|})$.
- Proof idea: Describe the run of the $po2dfa$ in $UITL$.
- Proof: Carry the context around. This is considerably more tedious than in the forward direction and yields a disjunction over many cases which in general gives an exponential upper bound for the size of the formula.

First order logic with two variables . . .

A context $u\underline{a}v$ can be written in $FO^2[S]$ as a formula.

For a formula $D_1 F_{u\underline{a}v} D_2$, the chop point is characterized by a formula $Mid_{\beta,\varepsilon}(y)$ which uses two formulae β and ε with one free variable as parameters standing for its beginning and ending points:

$$uav(y,z) \wedge \exists w \geq z : \varepsilon(w) \wedge \forall x < y(uav(x,z) \supset \exists y > x : \beta(y)).$$

w and z are metavariables standing for x or y, depending on whether the length of u, v is even or odd.

Getting there . . .

Translation A recursive version of the Gabbay trick of rotating variables:

$$
Tran_{\beta,\varepsilon}(D_1 F_{u\underline{a}v} D_2) = \begin{array}{l} \exists y \geq x(Mid_{\beta,\varepsilon}(y) \wedge Tran_{\beta(x),Mid_{\beta,\varepsilon}(y)}(D_1)) \wedge \\ \exists x \leq y(Mid_{\beta,\varepsilon}(x) \wedge Tran_{Mid_{\beta,\varepsilon}(x),\varepsilon(y)}(D_2)). \end{array}
$$

Finally, $\beta_0(x)$ and $\varepsilon_0(y)$ express away the two ends of a word.

. . . and coming back

Theorem An $FO^2[<, S]$-definable language is saturated by a Thérien-Wilke d, k-congruence for some $k > 0$, where $d - 1$ is the nesting depth of successor formulae.

Proof: Applying Thérien-Wilke on Rhodes expansions of words.

Construction of formulae For each d, k-equivalence class by induction on k and the d-content.

Summary

- Subclasses of starfree expressions with the same expressive power as the unambiguous languages of Schützenberger (1976), languages definable in $FO^2[<]$ and $FO^2[<, S]$.
- Automata for these classes with polynomial constructions.
- A “deterministic” logic admitting unique parsing of its models.
- A tractable fragment of ITL giving size $O(n^2)$ $po2dla$ construction and polynomial time modelchecking.
- Not succinct. To describe an automaton of n states requires an $O(2^n)$ length formula.

Summary and questions

TL	Satisfiability	Interval TL	Membership	Satisfiability
LTL[X,U]	Pspace	ITL	Pspace	Nonelementary
LTL[X,Y,F,P]	Pspace	LITL	LogDCFL	NP/Pspace
LTL[F,P]	NP	UITL	LogDCFL	NP

- Can we improve the succinctness?
- We have an NC^1 lower bound for membership (the same as for propositional logic). Can we narrow the gap?
- Is there a more expressive logic with NP satisfiability and Ptime membership?

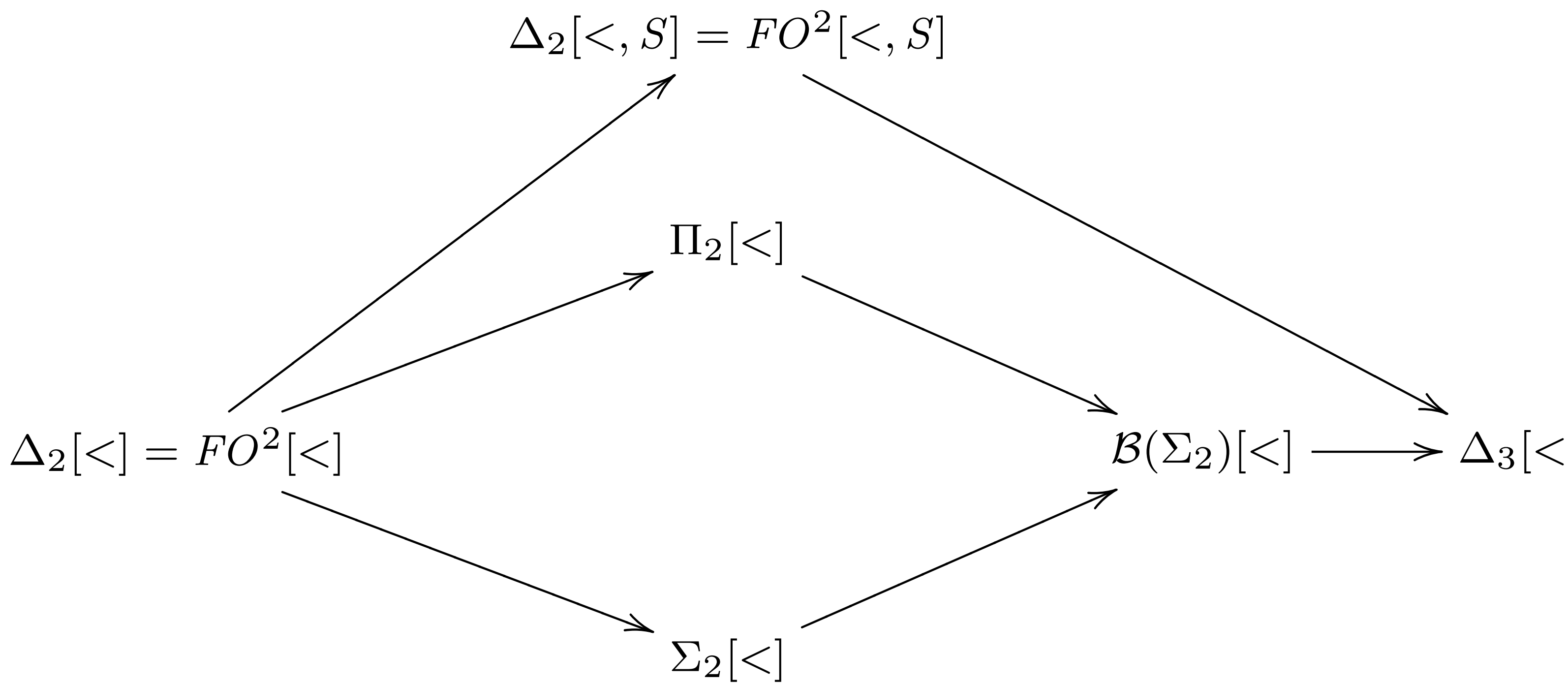
More open questions

Languages	Interval TL	Membership	Satisfiability
SF ⋮ Dotdepth 2	ITL ⋮ ?	Nonelementary ?	Nonelementary ?
UL	UITL	LogDCFL	NP/Pspace
Dotdepth 1	?	?	?

The dotdepth hierarchy of starfree expressions matches the quantifier alternation hierarchy in $FO[<]$.

- How far higher can we retain NP satisfiability and Ptime membership? Open.
- $FO^2[<]$ corresponds to Δ_2 in FO quantifier alternation between the first and second levels of dotdepth. What is the expressiveness of $FO^2[<, S]$? Answered.

Going around dot depth two



Theorem $(ac^*bc^*)^*$ is in $\Pi_2[<] \setminus FO^2[<, S]$

Theorem $\neg(\top L_{b\underline{b}} \top) F_{\underline{a}a} \top$ is in $FO^2[<, S] \setminus \mathcal{B}(\Sigma_2)[<]$

Proofs: Ehrenfeucht-Fraïssé games.

Acknowledgements

Project: The visit of Manfred Kufleitner to Chennai during Feb-Mar 2007 and of KL to Bordeaux during May-June 2007.

Discussions with Manfred and with Pascal Weil gave us a deeper understanding of the connection to the variety DA.

Pascal enlightened us on the distinction between unambiguous vs. deterministic and co-deterministic products.

Meena Mahajan sharpened our complexity arguments.

And many thanks for your attention.