

Zwei Reguläre Schritte für mich; ein großer für die Menschheit?

Zweistufige Ansätze für Grammatikformalismen

Frank Morawietz

<http://tcl.sfs.uni-tuebingen.de/~frank/>

Seminar für Sprachwissenschaft
Theoretische Computerlinguistik
Universität Tübingen

29. Januar 2002



- Einleitung und Motivation
- Reguläre Abfrage kontextsensitiver Relationen
 - Kontextfreie Baumgrammatiken
 - MSO-Abfragen
 - Lifting
 - Geliftete MSO-Abfragen
 - Rekonstruktion der intendierten Bäume
 - Intendierte Abfrageresultate
- Zusammenfassung und Ausblick





Einleitung und Motivation

- **Semistrukturierte Daten**

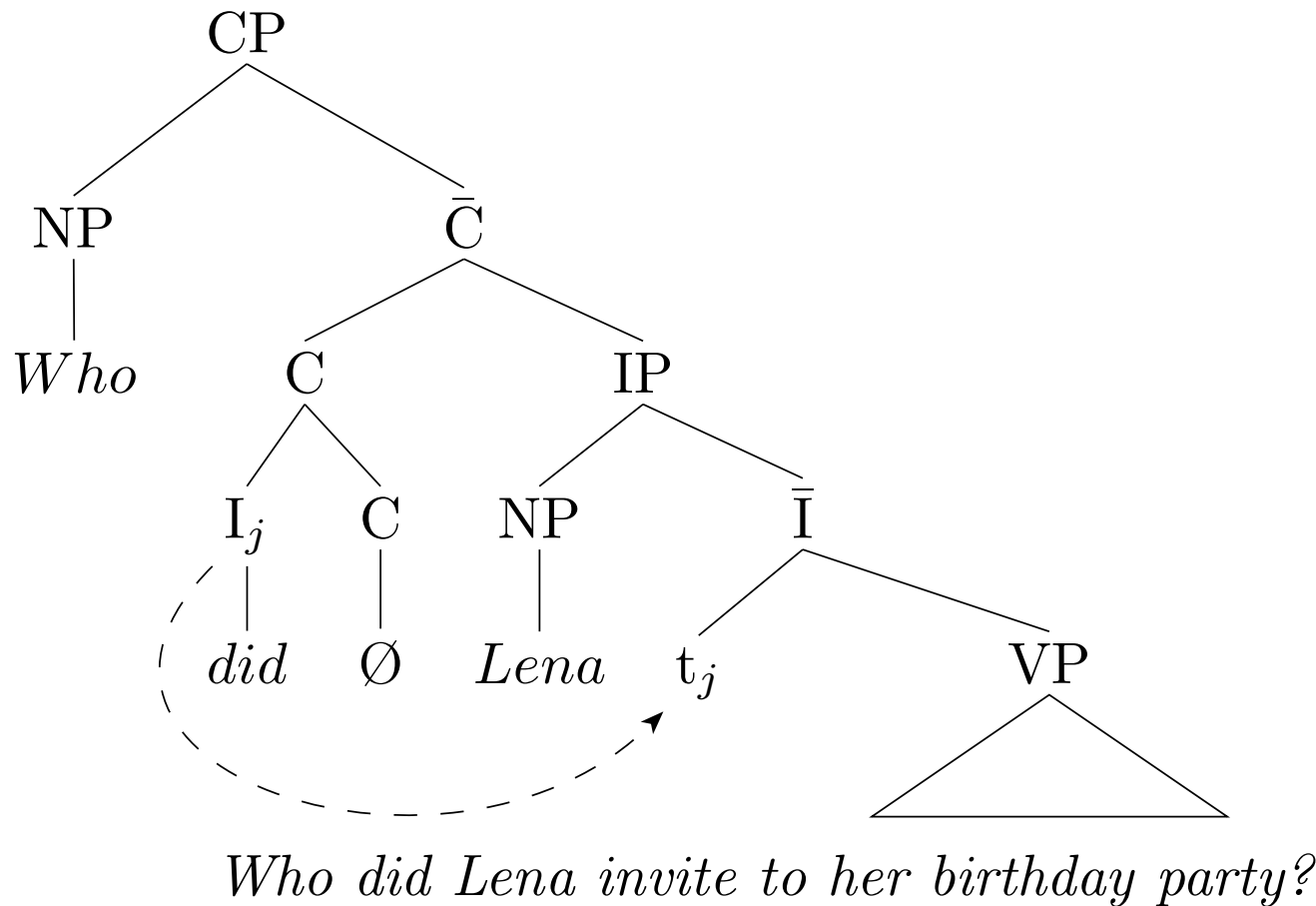
- Unvollständigkeit und Irregularität von Daten
- Integration verschiedener Daten
- Fehlen von relationalen Schemata

- **XML**

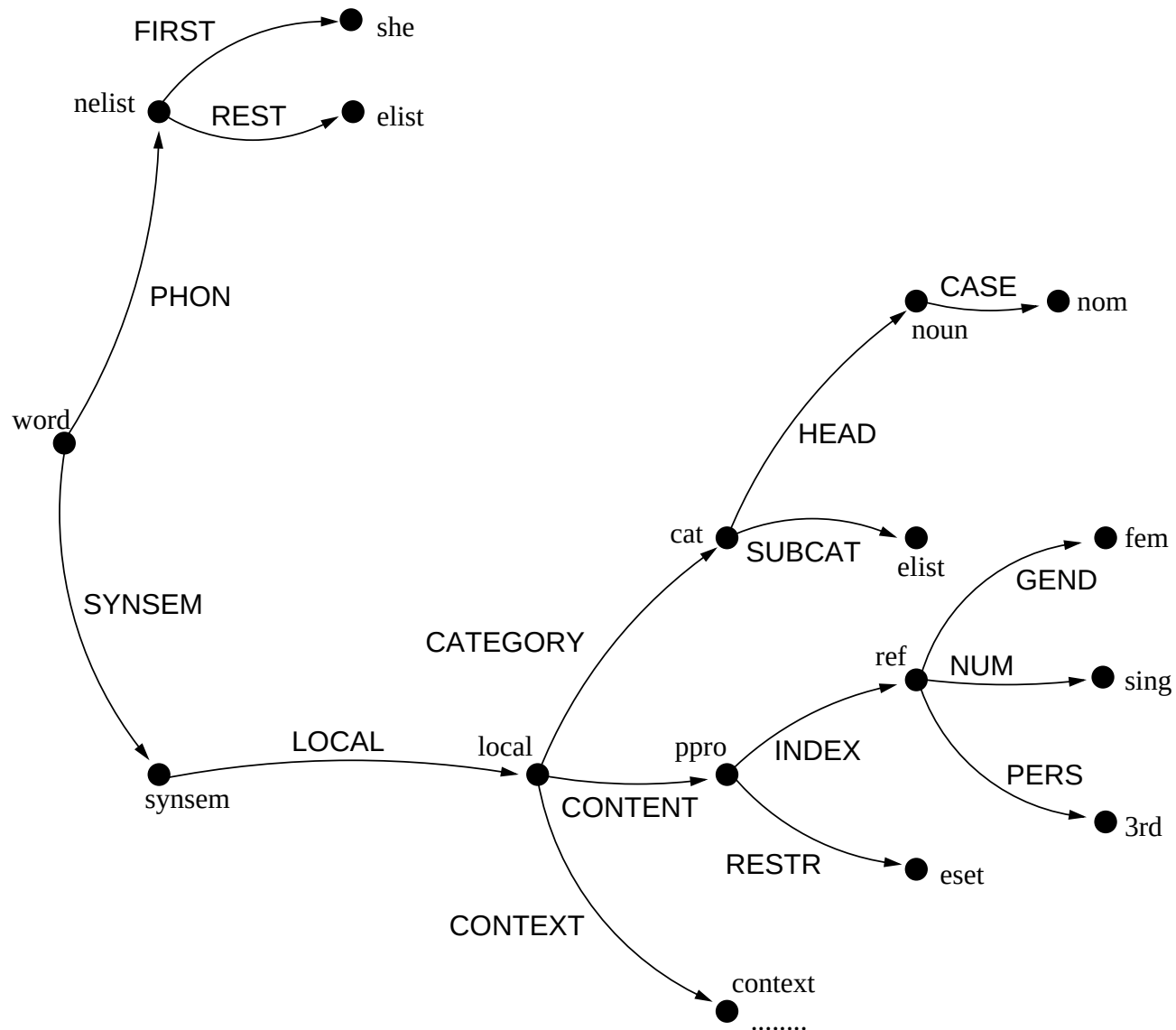
- Universelles Format des Datenaustauschs
- Abstraktes Datenmodell: Etikettierte Graphen
- Abfragesprachen für Datenmodell



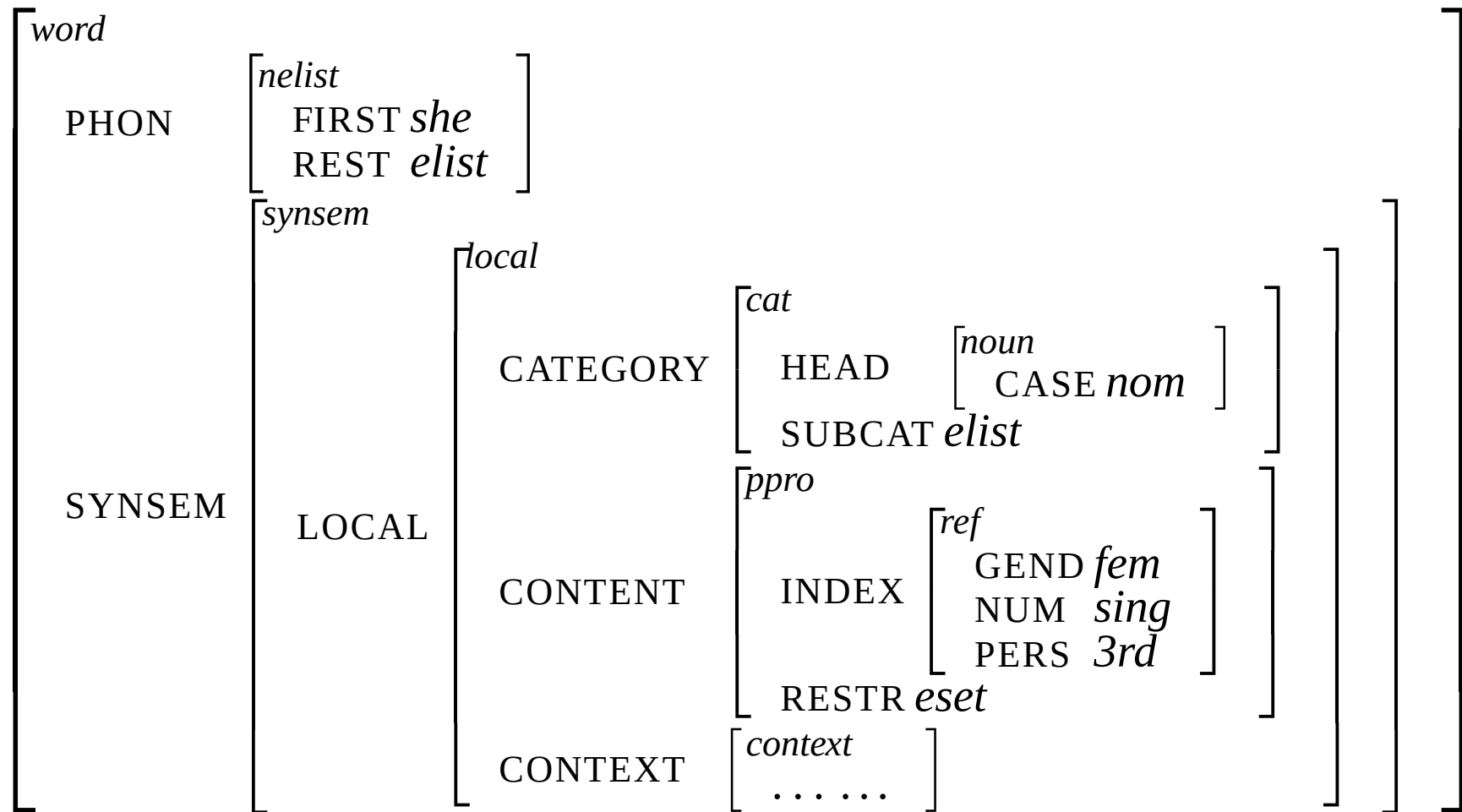
Semistrukturierte Daten in der Linguistik



Semistrukturierte Daten in der Linguistik



Semistrukturierte Daten in der Linguistik



Semistrukturierte Daten in der Linguistik

Penn Treebank II:

(*TOP* (*S* (*NP-SBJ* my best friend)
(*VP* gave
(*NP* me)
(*NP* chocolate)
(*NP-TMP* yesterday))
.))



Semistrukturierte Daten in der Linguistik

TGrep:

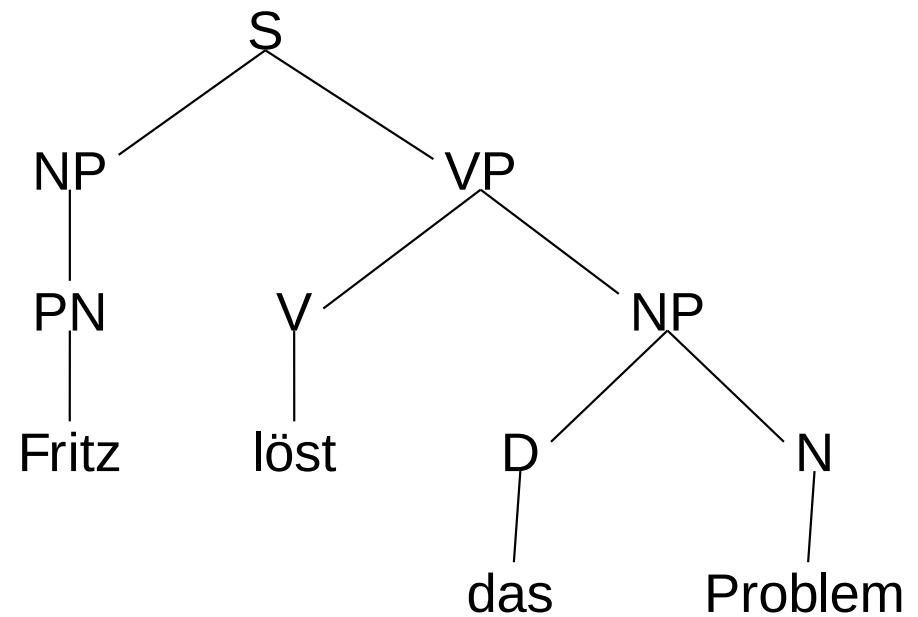
$$S <1 \ /^{NP}/ < (VP < (NP \$.. NP))$$

Finde alle *S* die mit einer *NP* beginnen (nicht unbedingt Subjekt), die eine *VP* dominieren, die zwei *NP* Töchter hat – mit anderen Worten, Sätze mit zwei Objekten.



Core XML

```
<S>
  <NP>
    <PN> Fritz </PN>
  </NP>
  <VP>
    <V> löst </V>
    <NP>
      <D> das </D>
      <N> Problem </N>
    </NP>
  </VP>
</S>
```



Fundamentales Problem

Skylla (Fehlende Ausdrucksstärke)

und

Charybdis (Unentscheidbarkeit)



Fundamentales Problem II

Klassisch

Semistrukturiert

RA

TAX

FO

MSO

SQL

Baumbeschreibungssprache
mit regulären Pfadausdrücken



Kreuzabhängigkeiten

*(... weil) der Karl die Maria dem Peter den Hans
schwimmen lehren helfen liess*

Deutsch: Dijck-Sprache—CF

*(... omdat) Karel Marie Piet Jan liet helpen leren
zwemmen*

Niederländisch: $a^n b^n$ —CF

(... wil) mer de maa em chind lönd hällfe schwüme



Kreuzabhängigkeiten

*(... weil) der Karl die Maria dem Peter den Hans
schwimmen lehren helfen liess*

Deutsch: Dijck-Sprache—CF

*(... omdat) Karel Marie Piet Jan liet helpen leren
zwemmen*

Niederländisch: $a^n b^n$ —CF

(... wil) mer de maa em chind lönd hällfe schwüme



Kreuzabhängigkeiten

*(... weil) der Karl die Maria dem Peter den Hans
schwimmen lehren helfen liess*

Deutsch: Dijck-Sprache—CF

*(... omdat) Karel Marie Piet Jan liet helpen leren
zwemmen*

Niederländisch: $a^n b^n$ —CF

(... wil) mer de maa em chind lönd hälffe schwüme



Kreuzabhängigkeiten

*(... weil) der Karl die Maria dem Peter den Hans
schwimmen lehren helfen liess*

Deutsch: Dijck-Sprache—CF

*(... omdat) Karel Marie Piet Jan liet helpen leren
zwemmen*

Niederländisch: $a^n b^n$ —CF

(... wil) mer de maa em chind lönd hällfe schwüme

Schweizerdeutsch: $a^n b^m c^n d^m$ —Nicht-CF



Reguläre Abfrage kontextsensitiver Relationen



Überblick: zweistufiger Ansatz

- Annahme: Es existiert eine Beschreibung der (mild) kontextsensitiven Relationen (z.B. durch eine kontextfreie Baumgrammatik)
- Datenbankabfrage
 - MSO Anfrage an die Baumbank
 - ⇒ Menge von möglichen Antworten/Bäumen
 - Lifting der Grammatik und der möglichen Antwortbäume
 - ⇒ Baumautomat für die Grammatik und eine geliftete Menge von möglichen Antwortbäumen
 - Automat filtert die gelifteten möglichen Antwortbäume
 - ⇒ Menge von gelifteten Antwortbäumen
- Rekonstruktion der intendierten Bäume durch eine MSO-Transduktion
 - ⇒ Menge der intendierten Antwortbäume



Kontextfreie Baumgrammatiken

Eine kontextfreie Baumgrammatik Γ ist ein 5-tuple $\langle \Sigma, \mathbf{F}, S, \mathbf{X}, \mathbf{P} \rangle$, mit

- $\Sigma, \mathbf{F}$ “gerankte” Alphabete von
Inoperativen und *Operativen*;
- $S \in \mathbf{F}$ Startsymbol;
- $\mathbf{X}$ eine Menge von Variablen; und
- $\mathbf{P}$ eine Menge von Produktionen.



Kontextfreie Baumgrammatiken

Die Produktionen $p \in \mathbf{P}$ haben die Form

$$F(x_1, \dots, x_n) \longrightarrow t$$

mit $F \in \mathbf{F}$, $x_1, \dots, x_n \in \mathbf{X}$, und t ein Term über $\Sigma \cup \mathbf{F} \cup \{x_1, \dots, x_n\}$.

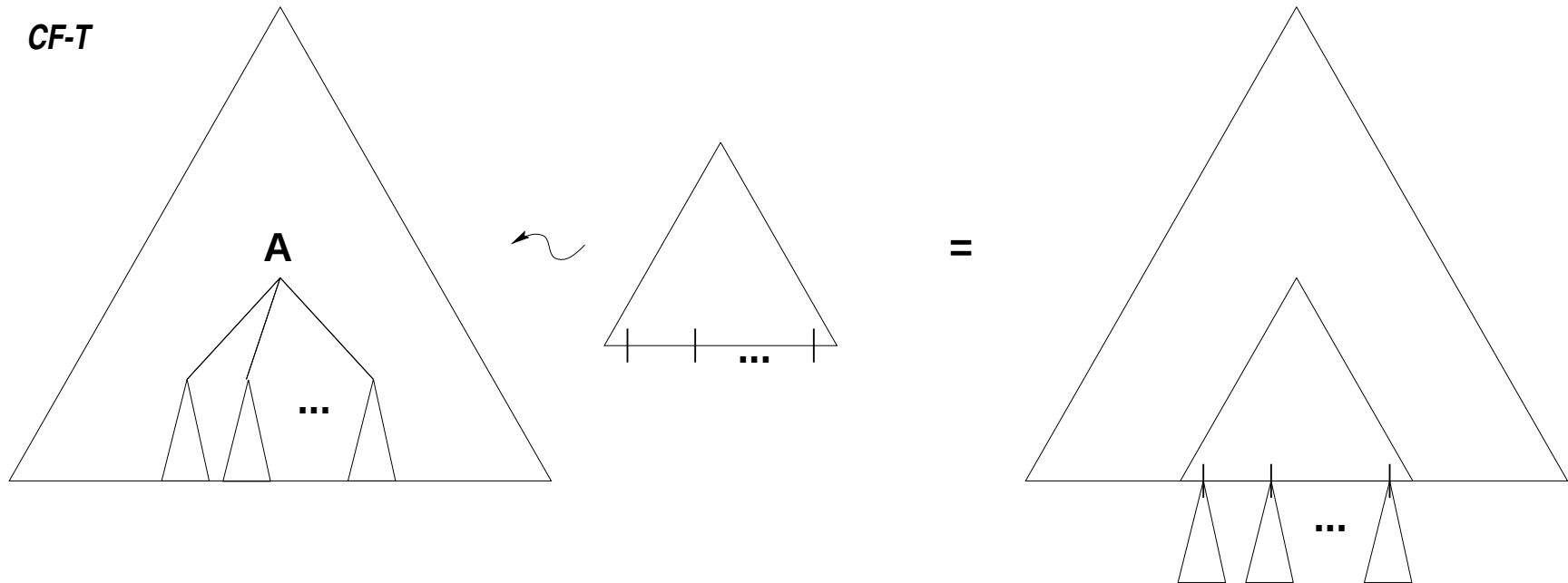
Eine Anwendung einer Regel $F(x_1, \dots, x_n) \longrightarrow t$ “ersetzt” einen Unterbaum, der in F verwurzelt ist, als den Baum t , wobei die Variablen durch die Töchter von F substituiert werden.

Baumgrammatiken für die $\mathbf{F}_n =$, $n \neq 0$, gilt, heißen regulär.

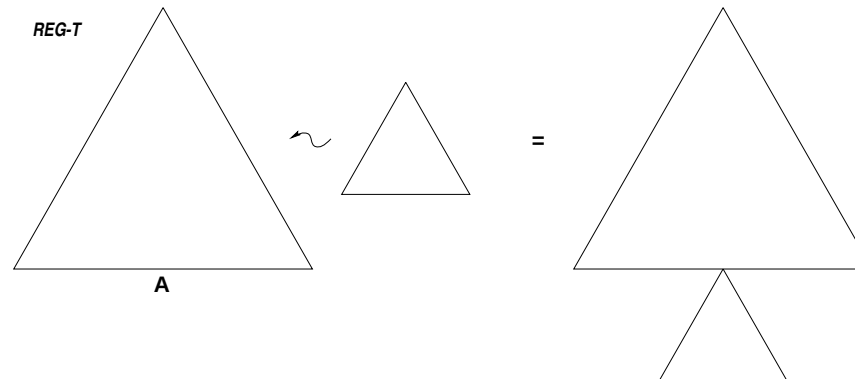


Kontextfreie Baumgrammatiken

CF-T



REG-T



Eine CFTG für $a^n b^m c^n d^m$: $\mathcal{G}$

$$\Sigma_0 = \{\varepsilon, a, b, c, d\}$$

$$\Sigma_2 = \{\bullet\}$$

$$\mathbf{F}_0 = \{S\}$$

$$\mathbf{F}_4 = \{F\}$$

$$S = S$$

$$\mathbf{X} = \{x_1, x_2, x_3, x_4\}$$

$$\mathbf{P} = \left\{ \begin{array}{ll} S & \longrightarrow \varepsilon \\ S & \longrightarrow F(a, \varepsilon, c, \varepsilon) \\ S & \longrightarrow F(\varepsilon, b, \varepsilon, d) \\ F(x_1, x_2, x_3, x_4) & \longrightarrow F(\bullet(a, x_1), x_2, \bullet(c, x_3), x_4) \\ F(x_1, x_2, x_3, x_4) & \longrightarrow F(x_1, \bullet(b, x_2), x_3, \bullet(d, x_4)) \\ F(x_1, x_2, x_3, x_4) & \longrightarrow \bullet(\bullet(\bullet(x_1, x_2), x_3), x_4) \end{array} \right\}$$

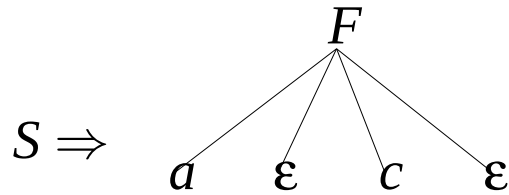


Eine Beispielableitung für $aabcccd$

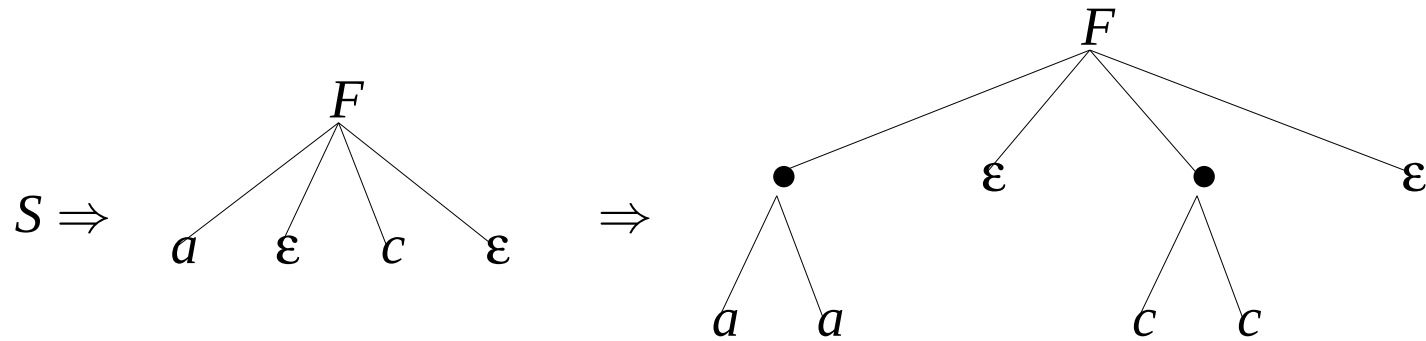
S



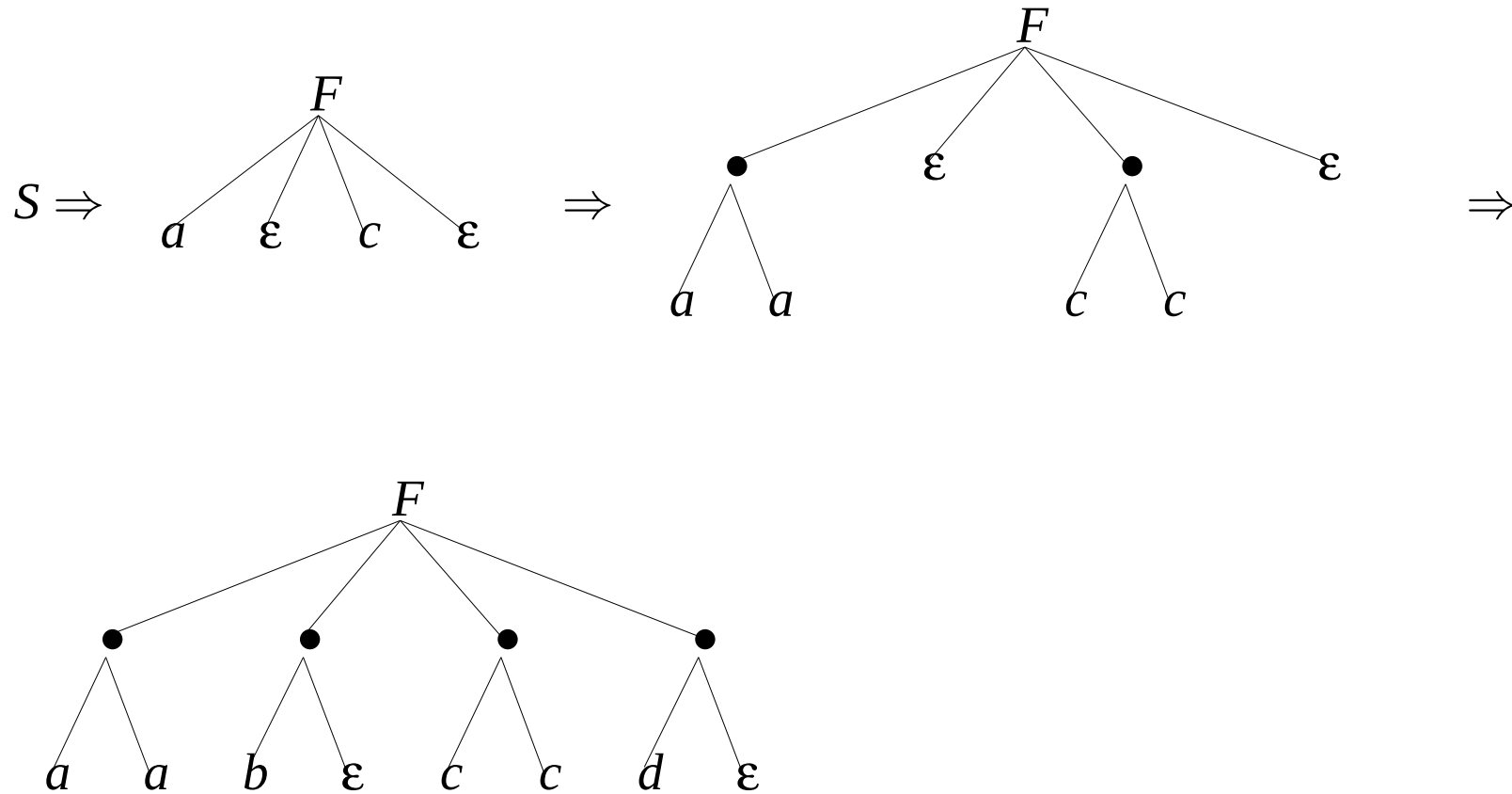
Eine Beispielableitung für $aabccd$



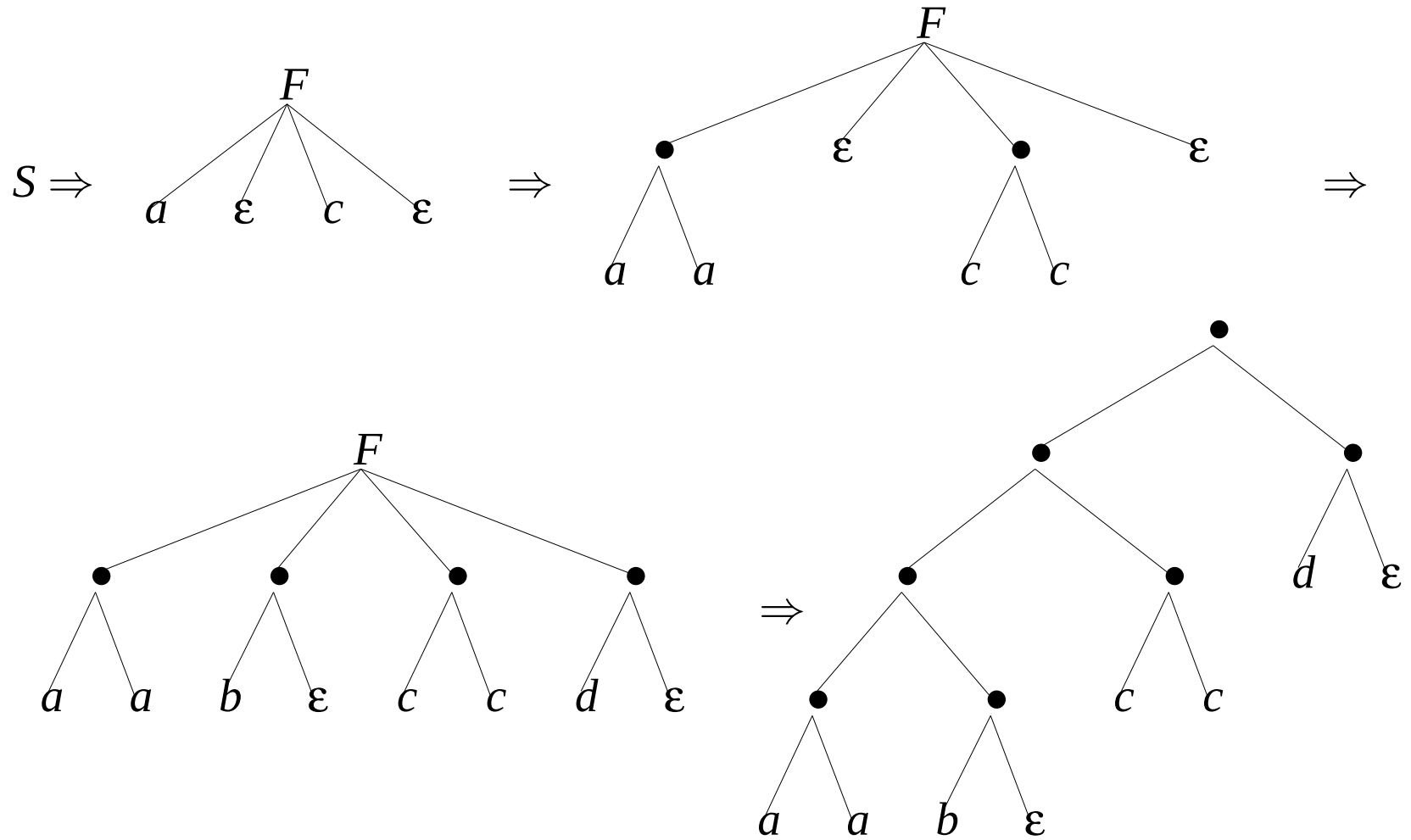
Eine Beispielableitung für $aabccd$



Eine Beispielableitung für $aabccd$



Eine Beispielableitung für $aabccd$



$$Q = \{t \models \varphi_r\}$$

$r \xLeftrightarrow{def}$ Regulärer Baumausdruck

$\varphi_r \xLeftrightarrow{def}$ Übersetzung von r in eine MSO Formel

Beispiel:

$$(NP_{akk} \cup NP_{dat})^* \cdot _ \cdot (V_{akk} \cup V_{dat})^*$$



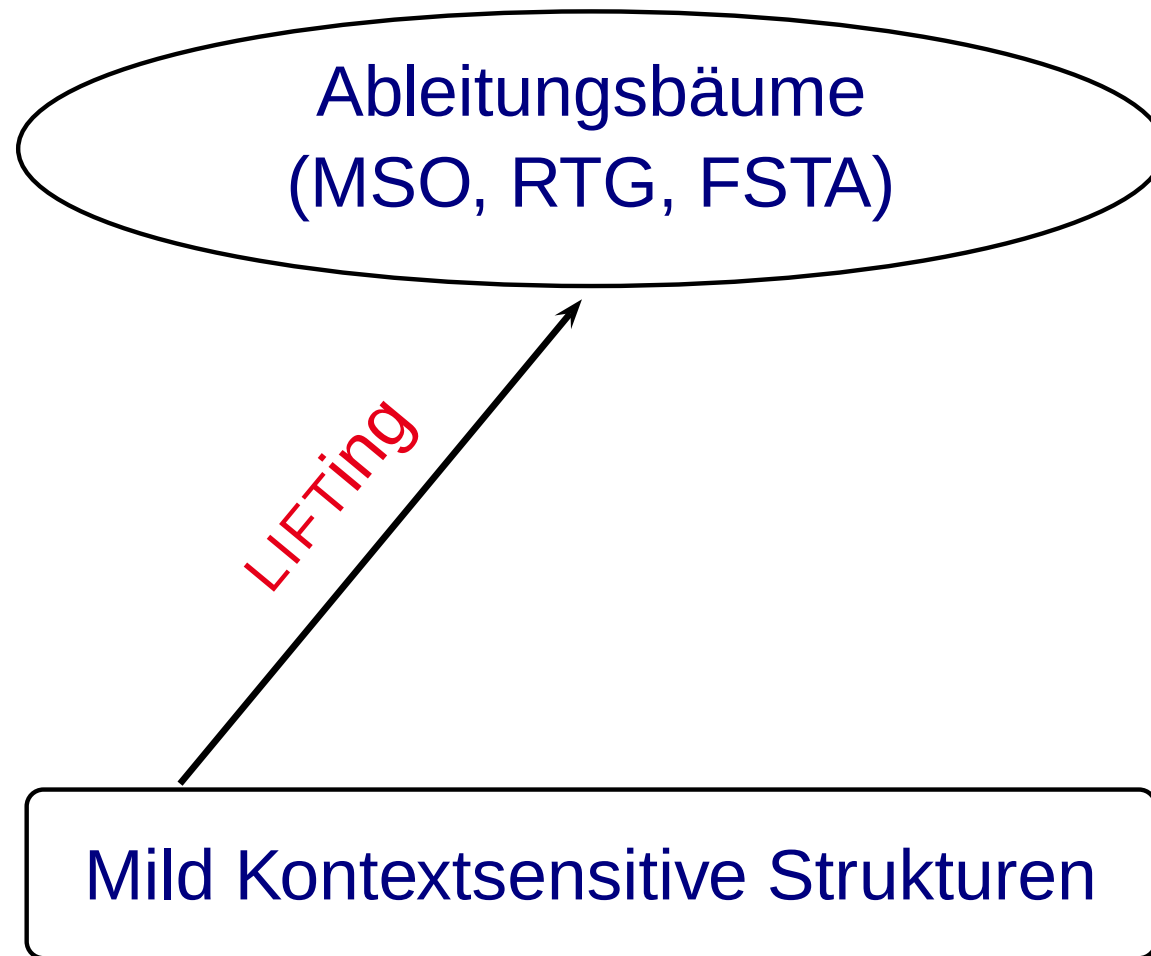
Überblick zwei Schritte

Ableitungsbäume
(MSO, RTG, FSTA)

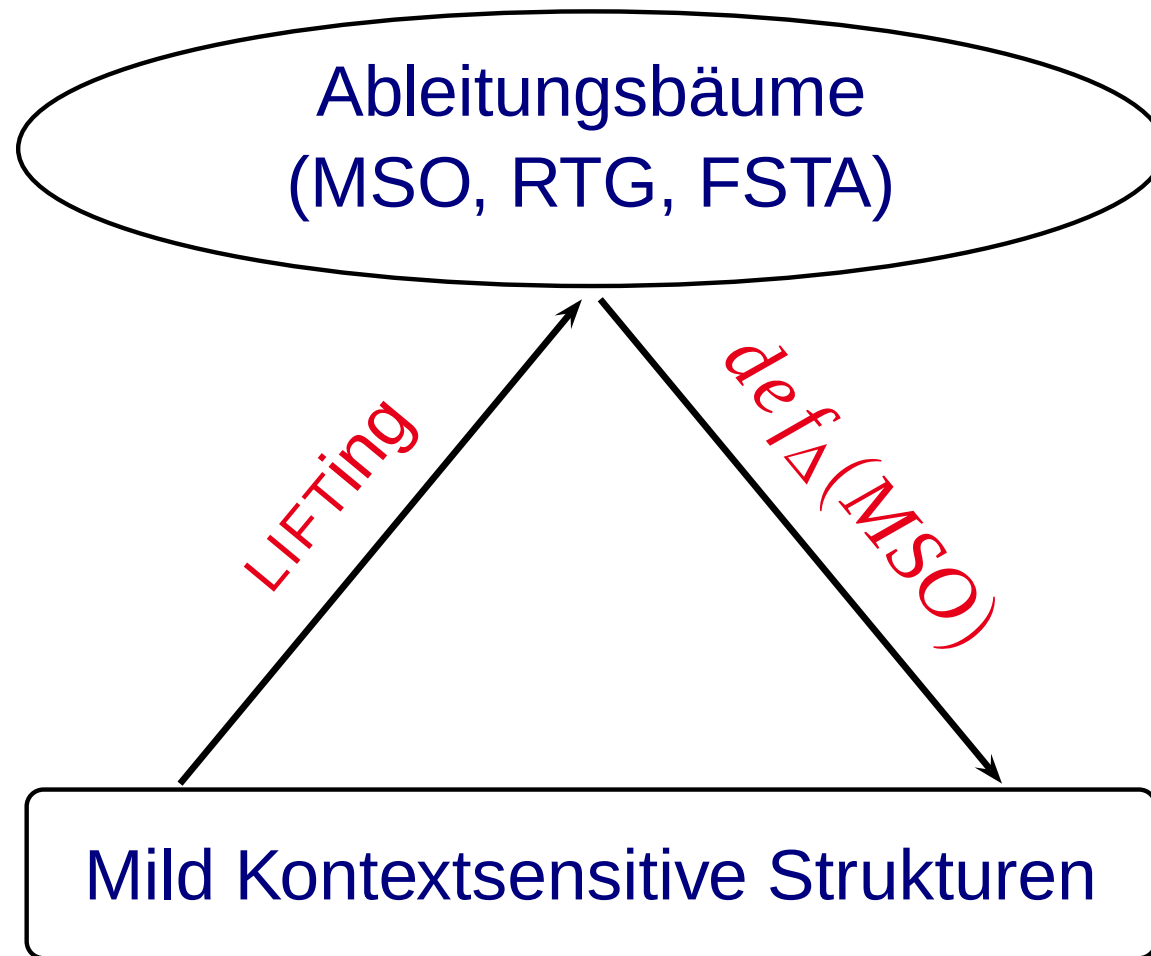
Mild Kontextsensitive Strukturen



Überblick zwei Schritte



Überblick zwei Schritte



Lifting: Intuition

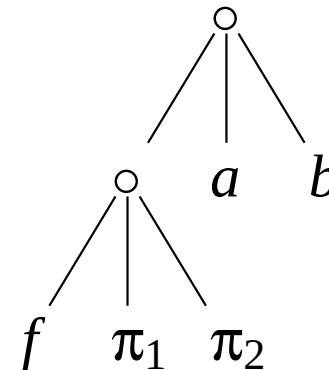
- Einfache primitiv-rekursive Funktion.
- Macht die Kontrollstruktur sichtbar.
- Alle Funktionssymbole werden zu Konstanten.
- Resultierende Grammatik ist *regulär*;
- und daher durch eine MSO-Formel ausdrückbar, und
- in einen Baumautomaten übersetzbar.



Lifting: Intuition

- Einfache primitiv-rekursive Funktion.
- Macht die Kontrollstruktur sichtbar.
- Alle Funktionssymbole werden zu Konstanten.
- Resultierende Grammatik ist *regulär*;
- und daher durch eine MSO-Formel ausdrückbar, und
- in einen Baumautomaten übersetzbar.

$$f(a, b) \rightsquigarrow (f \circ (\pi_1, \pi_2)) \circ (a, b) \rightsquigarrow$$



Für $k \geq 0$, $\text{LIFT}_k^\Sigma : T(\Sigma, \mathbf{X}_k) \rightarrow T(\Sigma^L, k)$ ist folgendermaßen definiert:

$$\text{LIFT}_k^\Sigma(x_i) = \pi_i^k$$

$$\text{LIFT}_k^\Sigma(f) = c_{(0,k)}(f') \text{ für } f \in \Sigma_0$$

$$\text{LIFT}_k^\Sigma(f(t_1, \dots, t_n)) =$$

$$c_{(n,k)}(f', \text{LIFT}_k^\Sigma(t_1), \dots, \text{LIFT}_k^\Sigma(t_n))$$

für $f \in \Sigma_n, n \geq 1$





Geliftete Abfrage

$$Q^L = \{t^L \models \varphi_r^L \wedge \psi\}$$

$$\varphi_r^L \stackrel{def}{\iff} \text{“Geliftetes” } \varphi_r$$

$$\psi \stackrel{def}{\iff} \text{“Geliftete” Spezifikation von} \\ \text{Kreuzabhängigkeiten}$$

Aber: die resultierenden Bäume sind nicht im Format der Baumbank und für Menschen/Linguisten ;-)
unlesbar.



Ziel der Rekonstruktion

- Der intendierte Baum ist im gelifteten Baum enthalten.
- Homomorphismus:

$$h(f') = f(x_1, \dots, x_n) \text{ für } f \in \Sigma_n$$

$$h(\pi_i^n) = x_i$$

$$h(c_{(n,k)}(t, t_1, \dots, t_n)) = h(t)[h(t_1), \dots, h(t_n)]$$

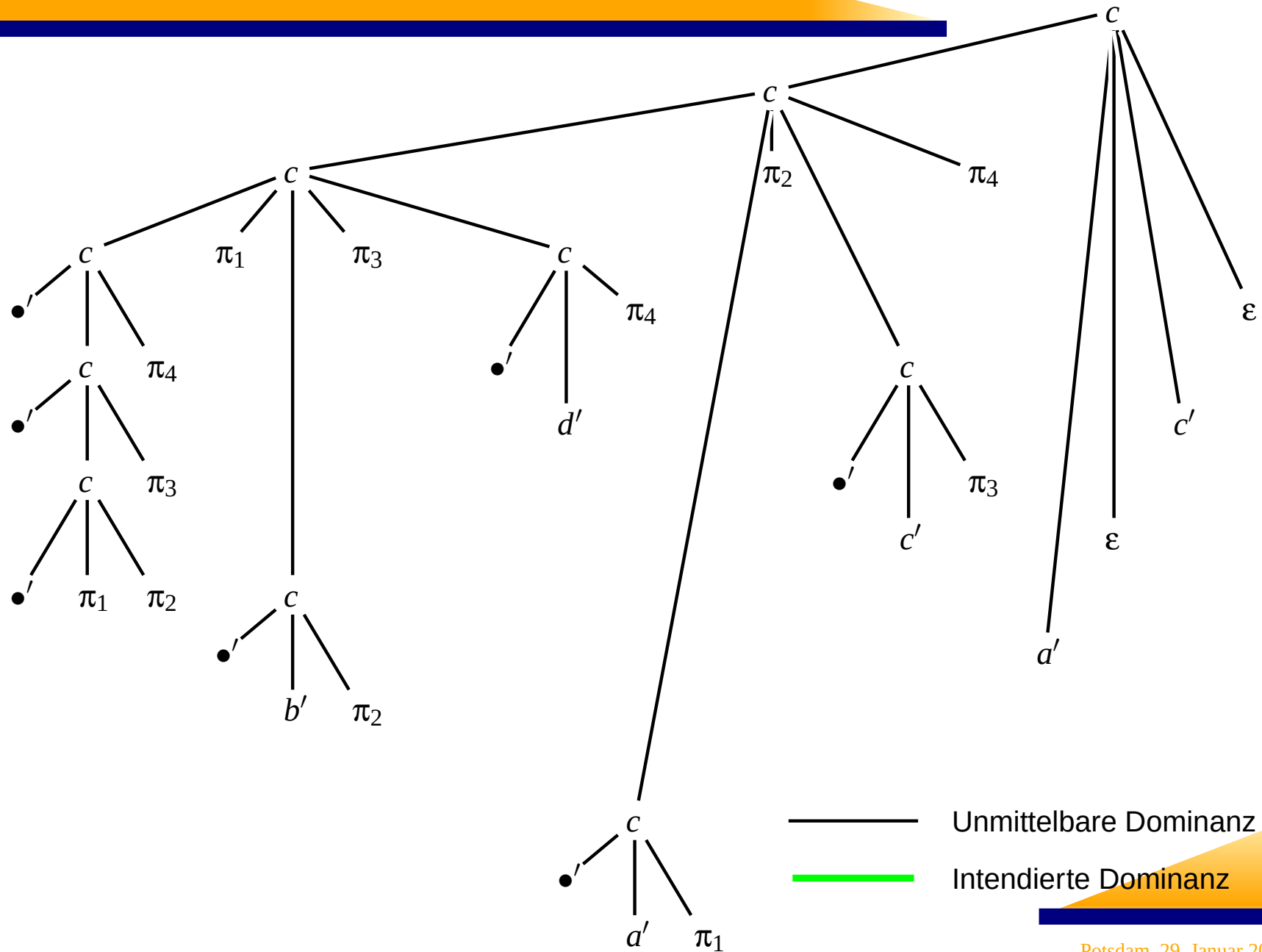
- Definition einer Menge von intendierten Relationen:

$$R^I = \{ \blacktriangleleft^* \text{ (Dominanz)}, \blacktriangleleft, \blacktriangleleft^+, \triangleleft \text{ (Präzedenz)}, \dots \}$$

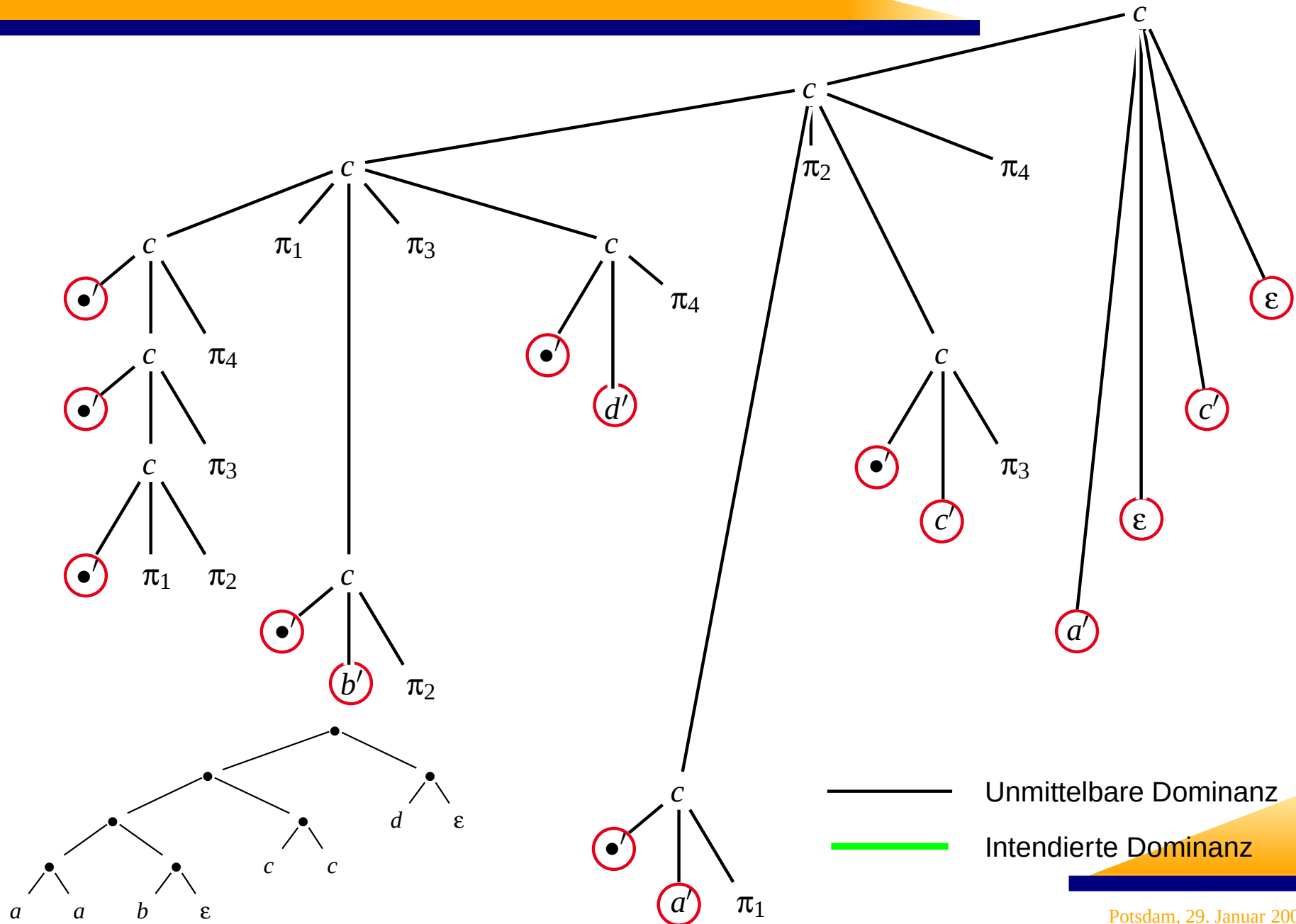
⇒ MSO-Transduktion



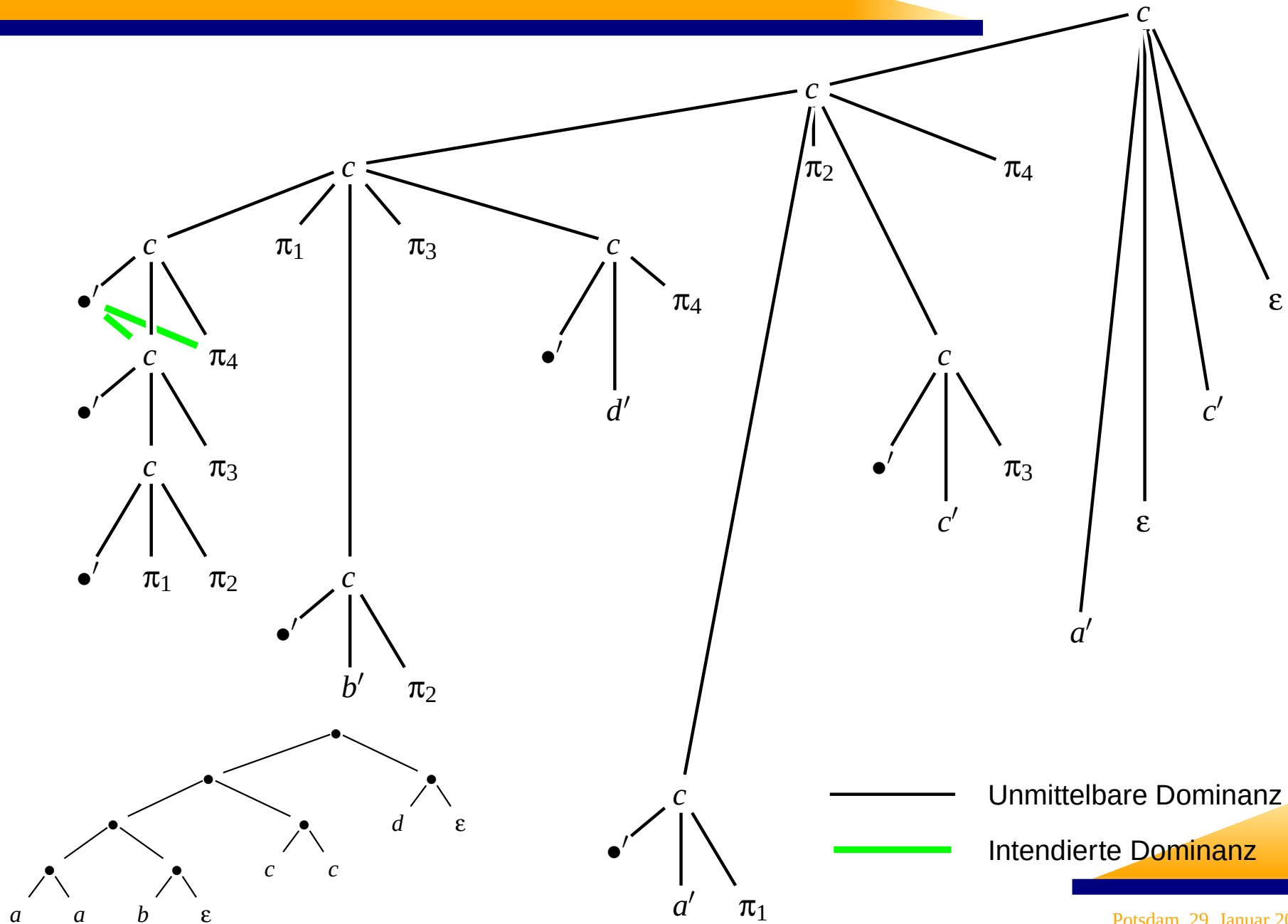
Rekonstruktion der Intendierten Strukturen



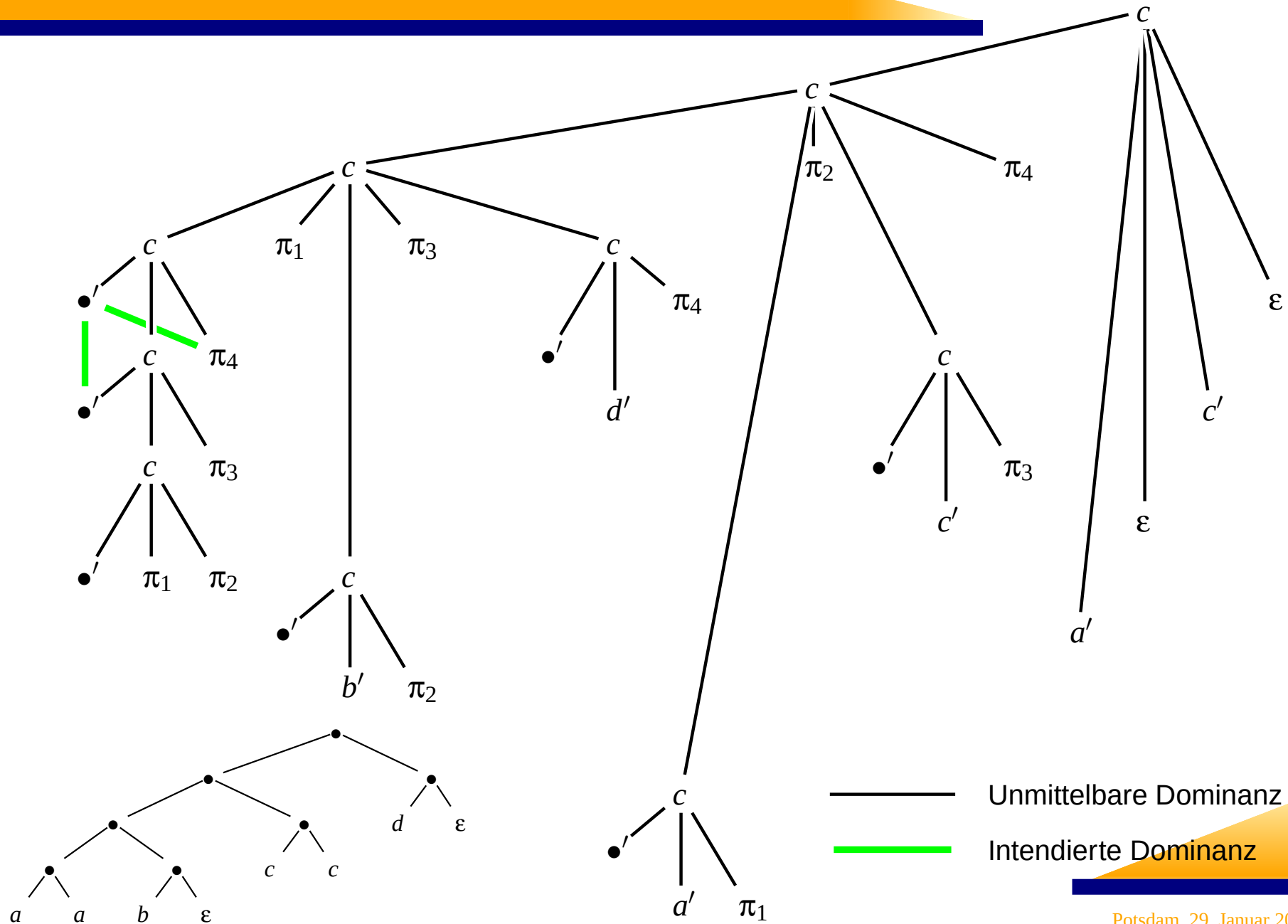
Rekonstruktion der Intendierten Strukturen



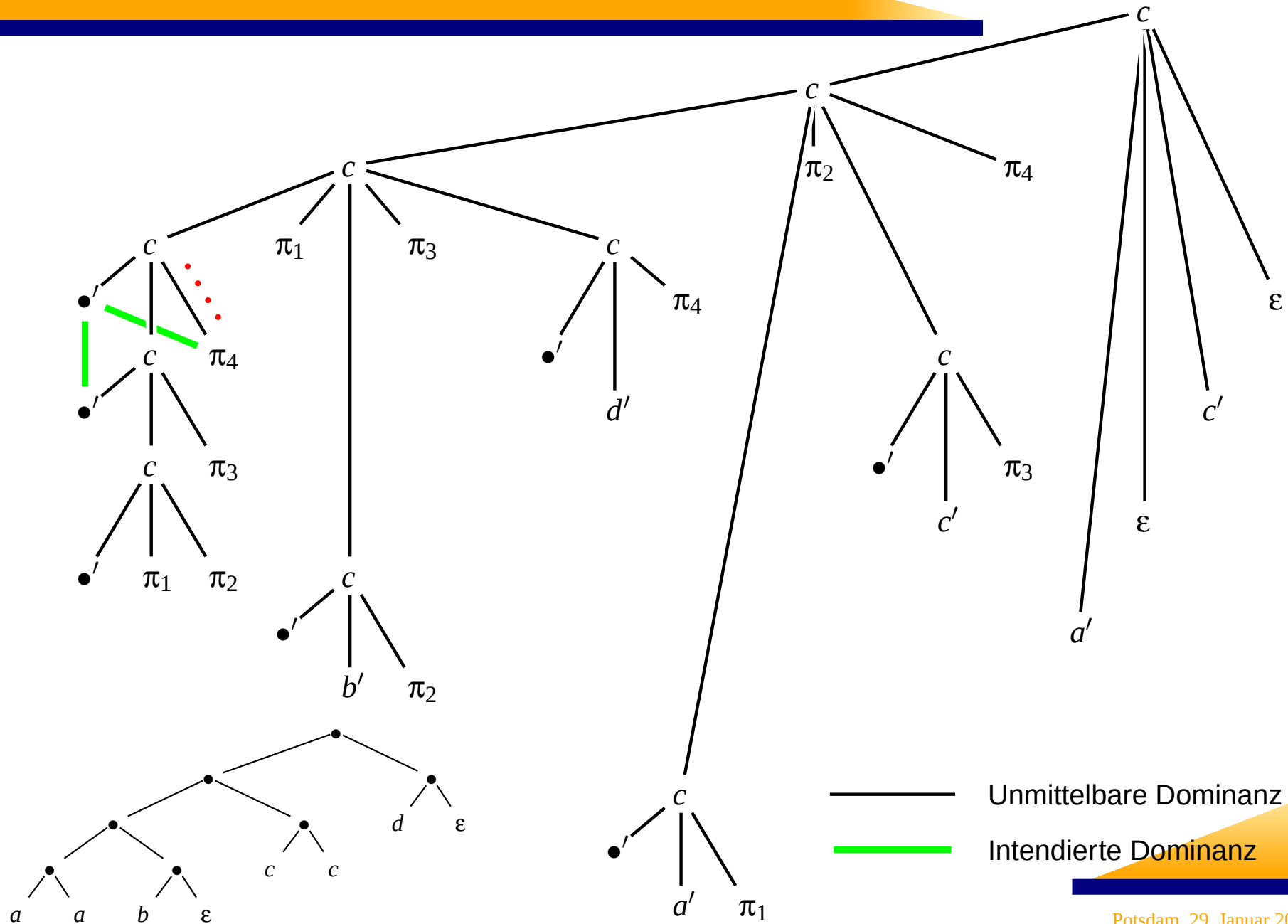
Rekonstruktion der Intendierten Strukturen



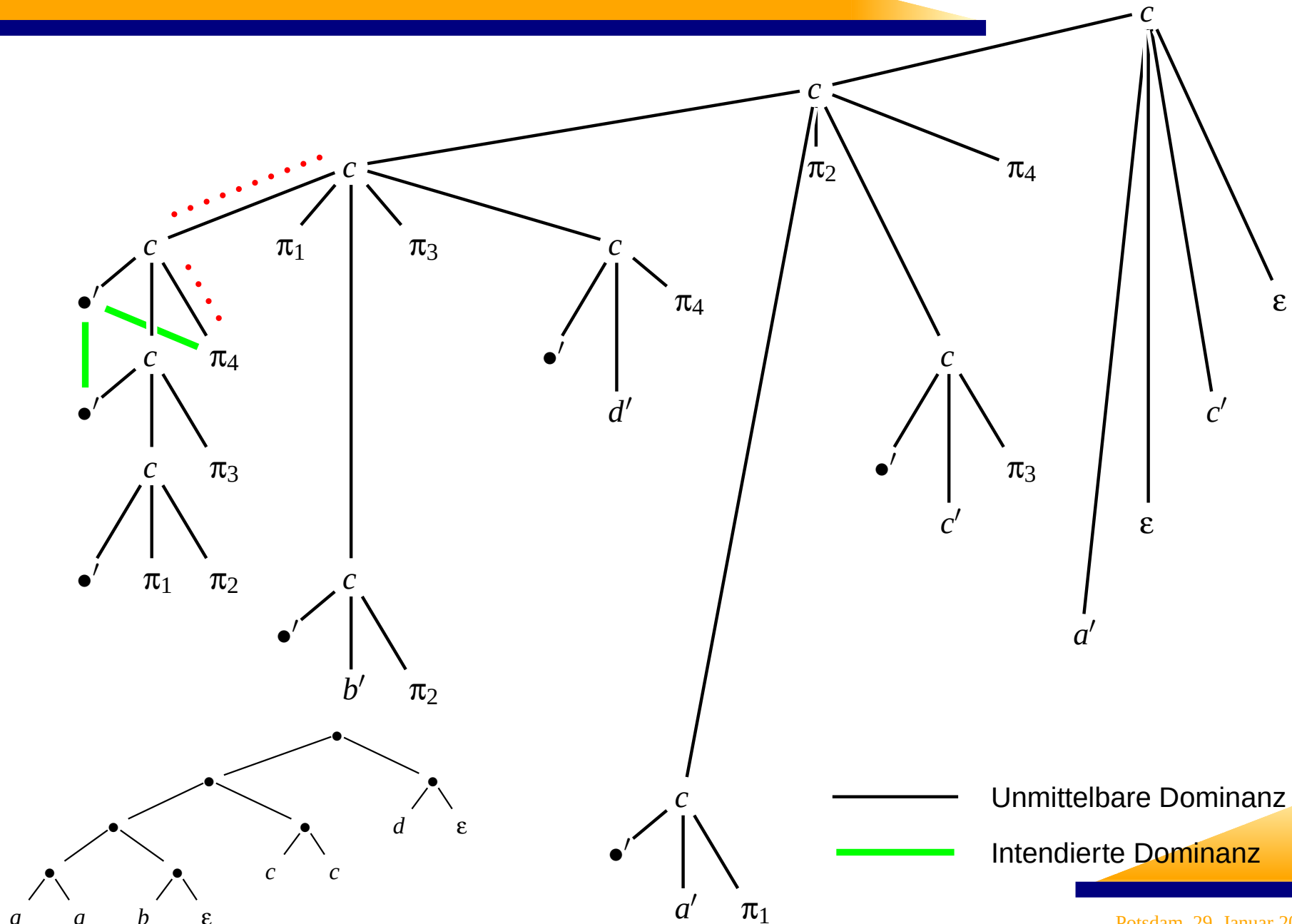
Rekonstruktion der Intendierten Strukturen



Rekonstruktion der Intendierten Strukturen



Rekonstruktion der Intendierten Strukturen



EBERHARD KARLS UNIVERSITÄT TÜBINGEN





EBERHARD KARLS UNIVERSITÄT TÜBINGEN





MSO definierbare Transduktionen

$$\mathcal{R} \rightsquigarrow \mathcal{S}$$

$$\Delta = (\chi, \xi, (\theta_s)_{s \in \mathcal{S}})$$

- χ der Definitionsbereich der Transduktion
- ξ der Bereich der (Ziel-) Struktur $\mathcal{S}$
- θ_s die neuen Relationen

Definition von Courcelle 1997



Intended Structures via MSO transduction

$$(\varphi, \psi, (\theta_s)_{s \in S})$$

$$S = \{\triangleleft, \triangleleft^*, \triangleleft^+, \triangleleft, \dots\}$$

$$\chi = \varphi_r^L \wedge \psi$$

ξ = Knoten mit einem “linguistischen” Label

$$\theta_{\triangleleft} = \text{trans}_{W_{\triangleleft}}(x, y)$$

$$\theta_{\triangleleft^*} = (\forall X)[\triangleleft\text{-closed}(X) \wedge x \in X \rightarrow y \in X]$$

$$\theta_{\triangleleft^+} = x \triangleleft^* y \vee x \not\approx y$$

$$\theta_{\triangleleft} = \text{trans}_{W_{\triangleleft}}(x, y)$$

$$\theta_{\text{label}} = \text{übernommen aus } R$$



Intendierte Abfrage-Resultate

$$Q' = \{def_{\Delta}(t^L) \mid t^L \models \phi_r^L \wedge \psi\}$$

$def_{\Delta} \stackrel{def}{\iff}$ MSO Transduktion vom Bereich der “gelifteten”
regulären Bäume in die ursprüngliche
Baumdatenbank

$\Delta \stackrel{def}{\iff}$ MSO definierbare Transduktion





Zusammenfassung und Ausblick

- Abfrage von nicht kontextfreien Relationen mit regulären Mitteln.
- drei “Arten” von Bäumen
 - Bäume als Elemente einer freien Algebra
 - Bäume als Elemente eines freien Klons (Lawvere-Algebra)
 - Bäume als relationale Strukturen
- Transduktion auch mit Hilfe von Makro-Baumtransduktoren



- Anpassung an ungerankte Bäume
- Implementation
- Effizienz
 - MSO auf Bäumen: entscheidbar
 - MSO auf Graphen: unentscheidbar
 - MSO auf Bäumen + eine binäre Relation: unentscheidbar
 - Übersetzung MSO $\rightarrow$ Baumautomaten: hyperexponentiell
 - “Guarded MSO” $\rightarrow$ Baumautomaten: exponentiell
 - Lifting und Transduktion: linear
- Lifting der möglichen Antwortbäume



- Einleitung und Motivation
- Reguläre Abfrage kontextsensitiver Relationen
 - Kontextfreie Baumgrammatiken
 - MSO-Abfragen
 - Lifting
 - Geliftete MSO-Abfragen
 - Rekonstruktion der intendierten Bäume
 - Intendierte Abfrageresultate
- Zusammenfassung und Ausblick





Appendix

Einige Details und Definitionen

Coding RTGs in MSO Logic (Thomas 1997)

- Define a tree automaton $\mathfrak{A}_{\mathcal{G}'}$ from the RTG $\mathcal{G}'$
- Code its behaviour in $\phi_{\mathfrak{A}_{\mathcal{G}'}}$

$$(\exists X_0, \dots, X_m) \left[\bigwedge_{i \neq j} (\neg \exists y) [y \in X_i \wedge y \in X_j] \wedge \right.$$

$$(\forall x) [(\neg \exists y) [x \triangleleft y] \rightarrow x \in X_0] \wedge \quad \% \text{ Initial State}$$

$$\bigwedge_{1 \leq n \leq m} (\forall x_1, \dots, x_n, y) \left[\bigvee_{\substack{(i_1, \dots, i_n, \sigma, j) \in \alpha \\ 1 \leq k \leq n}} x_k \in X_{i_k} \wedge y \triangleleft x_k \wedge y \in X_j \wedge y \in P_\sigma \right]$$

$$\bigvee_{i \in F} (\exists x \forall y) [x \triangleleft^* y \wedge x \in X_i] \quad \% \text{ Root}$$



Sei Σ ein ranked Alphabet. Das *abgeleitete* $\mathbb{N}$ -sortierte Alphabet Σ^L ist wie folgt definiert:

Für jedes $n \geq 0$,

$$\Sigma'_n = \{f' \mid f \in \Sigma_n\}$$

ist eine neue Menge von Symbolen der Sorte n ;

zurück zu Lifting



Sei Σ ein ranked Alphabet. Das *abgeleitete* $\mathbb{N}$ -sortierte Alphabet Σ^L ist wie folgt definiert:

für jedes $n \geq 1$ und jedes $i, 1 \leq i \leq n$,

$$\pi_i^n$$

ist ein neues Symbol, das *ite* Projektions-Symbol der Sorte n ;

zurück zu Lifting



Sei Σ ein ranked Alphabet. Das *abgeleitete* $\mathbb{N}$ -sortierte Alphabet Σ^L ist wie folgt definiert:

für jedes $n \geq 0, k \geq 0$ das neue Symbol

$$c_{(n,k)}$$

ist das (n, k) te Kompositions-Symbol.

zurück zu Lifting



Sei Σ ein ranked Alphabet. Das *abgeleitete* $\mathbb{N}$ -sortierte Alphabet Σ^L ist wie folgt definiert:

$$\Sigma_0^L = \Sigma'_0$$

$$\Sigma_n^L = \Sigma'_n \cup \{\pi_i^n \mid 1 \leq i \leq n\} \text{ für } n \geq 1$$

$$\Sigma_{n,k}^L = \{c_{(n,k)}\} \text{ für } n, k \geq 0$$

$$\Sigma_i^L = \text{sonst}$$

[zurück zu Lifting](#)



Model-Theoretic Interpretation

Basic Idea (Rabin 1965)

Obtaining a structure $\mathbb{B} = \langle B, Q \rangle$ from a structure $\mathbb{A} = \langle A, \mathcal{R} \rangle$ where $\mathcal{R}$ and Q are families of relation symbols.

zurück zu MSO-Transduktion



Tree-Walking Automaton

A *tree-walking automaton (with tests)* (Bloem and Engelfriet 1997) over some ranked alphabet Σ is a finite automaton $\mathfrak{A} = (Q, \Delta, \delta, I, F)$ with states Q , directives Δ , transitions $\delta : Q \times \Delta \rightarrow Q$ and the initial and final states $I \subseteq Q$ and $F \subseteq Q$ which traverses a tree using three kinds of directives:

$\uparrow_i$ “move up to the mother of the current node (if it has one and it is its i -th daughter)”,

$\downarrow_i$ “move to the i -th daughter of the current node (if it exists)”,

$\phi(x)$ “verify that ϕ holds at the current node”.



Regular Tree-Node Relations

For any tree t , such a tree-walking automaton $\mathfrak{A}$ computes a node relation

$$R_t(\mathfrak{A}) = \{(x, y) \mid (x, q_i) \xRightarrow{*} (y, q_f) \text{ for} \\ \text{some } q_i \in I \text{ and some } q_f \in F\}$$

where for all states $q_i, q_j \in Q$ and nodes x, y in t

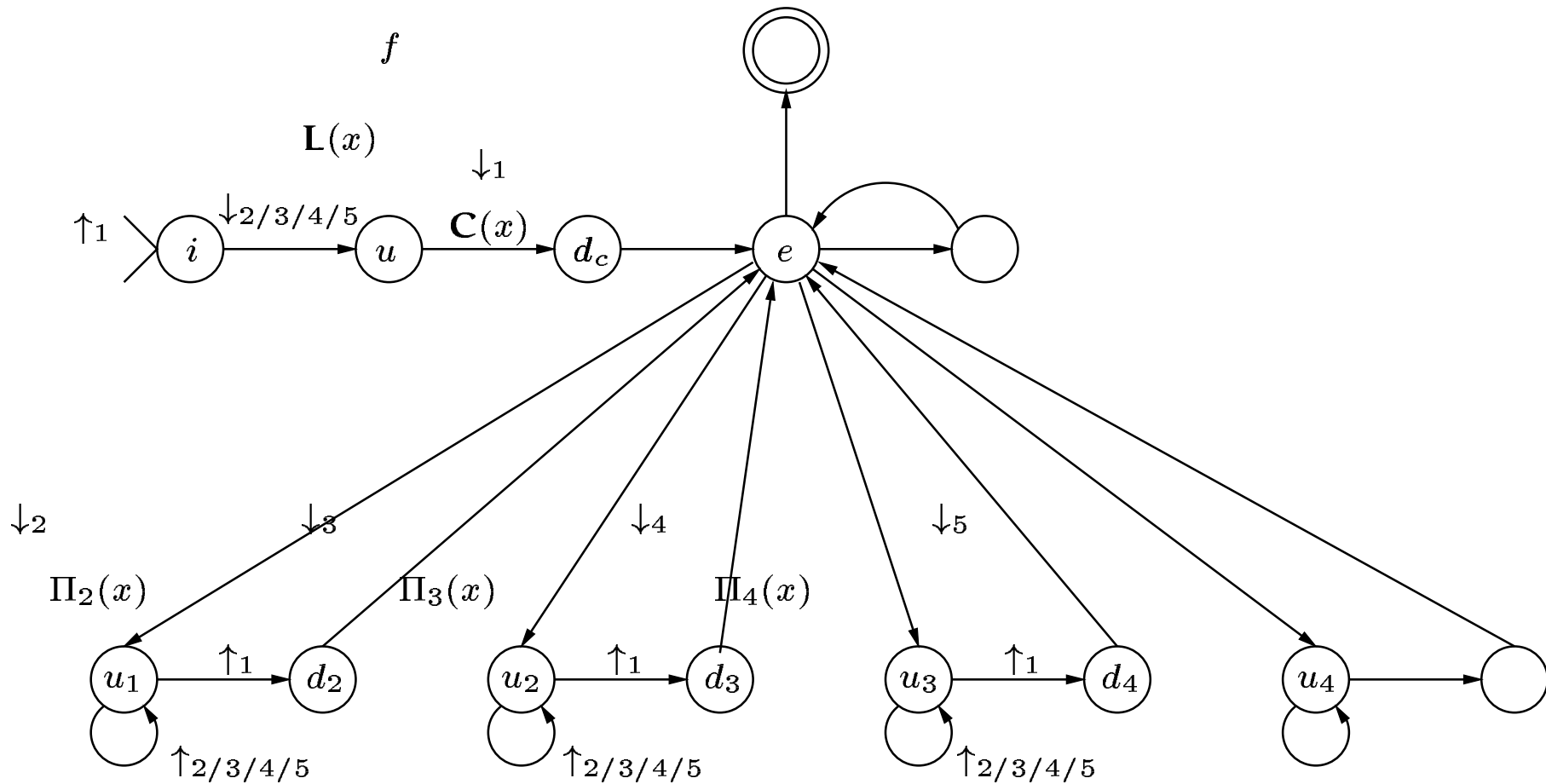
$$(x, q_i) \Rightarrow (y, q_j) \text{ iff } \exists d \in \Delta : (q_i, d, q_j) \in \delta \\ \text{and } y \text{ is reachable from } x \text{ in } t \text{ via } d.$$

If all the tests $\phi(x)$ of $\mathfrak{A}$ are *MSO* definable, $\mathfrak{A}$ specifies a *regular tree-node relation*, which is itself *MSO* definable.

[zurück](#) zu MSO-Transduktion



Der tree-walking Automat $\mathcal{A}$



zurück zu MSO-Transduktion



Reflexiver Abschluß in *MSO*

$$R\text{-closed}(X) \stackrel{\text{def}}{\iff} (\forall x, y)[x \in X \wedge R(x, y) \rightarrow y \in X]$$

$$R^*(x, y) \stackrel{\text{def}}{\iff} (\forall X)[R\text{-closed}(X) \wedge x \in X \rightarrow y \in X]$$

zurück zu MSO-Transduktion



Walking language $\Rightarrow$ MSO-formula

$$\text{trans}(x, y) = \perp$$

$$\text{trans}_{\downarrow_i}(x, y) = \text{edge}_i(x, y)$$

$$\text{trans}_{\uparrow_i}(x, y) = \text{edge}_i(y, x)$$

$$\text{trans}_{\phi(x)}(x, y) = \phi(x) \wedge x = y$$

$$\text{trans}_{W_1 \cup W_2}(x, y) = \text{trans}_{W_1}(x, y) \vee \text{trans}_{W_2}(x, y)$$

$$\text{trans}_{W_1 \cdot W_2}(x, y) = \exists z (\text{trans}_{W_1}(x, z) \wedge \text{trans}_{W_2}(z, y))$$

$$\text{trans}_{W^*}(x, y) = \text{trans}_W^*(x, y)$$

$$\begin{aligned} \text{trans}_W^*(x, y) = \forall X (\forall v, w (v \in X \wedge \text{trans}_W(v, w) \rightarrow w \in X) \\ \wedge x \in X \rightarrow y \in X) \end{aligned}$$

