Aspekte formaler Linguistik

Uwe Mönnich

Seminar für Sprachwissenschaft
Universität Tübingen

um@sfs.nphil.uni-tuebingen.de

http://tcl.sfs.nphil.uni-tuebingen.de/

Cusanus Stiftung, Mai 2000

Übersicht

I. Formale Datentypen zur Struktur natürlicher Sprachen und zur menschlicher Sprachfähigkeit

1. Algebraische und quantitative Methoden
2. Der Kognitive Ansatz
3. Montague–Grammatik
4. Paradoxien und Puzzles in der analytischen Sprachphilosophie
5. Die Chomsky–Hierarchie

II. Reguläre Beschreibung von Kreuzabhängigkeiten (zusammen mit H.P. Kolb und F. Morawietz)

1. Motivation
2. Kontextfreie Grammatiken
3. Die LIFT-Operation
4. Kodierung geLIFTeter kontextfreier Baumsprachen
5. Wiedergewinnung indentierter Strukturen
6. Kodierung indentierter Strukturen
7. Envoi

I.2. Der Kognitive Ansatz

In Descartes and Plato, as in Chomsky, the nativism is so striking that one is likely to overlook a still deeper consensus: the idea that certain of the subject's cognitive capacities should be explained by reference to consequence relations (e.g., deductive relations) that hold among the propositions that the subject knows (believes, cognizes, or whatever). I say to you: "What's 2 plus 17?" and you, being good at that sort of thing say "19." Your behavior *is structured in the relevant sense; what sort of* mental *structure is the psychologist to posit in explaining your behavior? According to the Cartesian, it is inter alia the* deductive *structure of number theory to which the explanation must appeal. (…) Similarly, according to generative linguistic theory, your ability to detect syntactic ambiguities, distinguish well-formedness from ungrammaticality, respond selectively to the noun-phrase that has been topicalized, and so forth are to be explained by reference to* what is entailed *by the grammar that you learned when you learned your language. In short, your linguistic capacities explain your verbal behavior, and are themselves explained by reference to the* content of your beliefs*. You can spot the ambiguity of "they are flying planes" because, so the story goes (i) You have learned the grammar of English, and (ii) it follows - deductively - from what you have learned that "they are flying planes" has two well-formed parsings.*
(Jerry Fodor: The Modularity of Mind, Cambridge, MA 1983, 6s)

Das Spektrum der verschiedenen Arten, in denen Computer in der Linguistik genutzt werden, ist breit. An seinem Ende steht der rein instrumentelle Einsatz zur Verarbeitung sprachlicher Daten. (. . .) Das andere Ende des Spektrums wird markiert durch das Konzept der Künstlichen Intelligenz in der sogenannten starken Version, die sich an der strengen Deutung der Computermetapher orientiert, daß nämlich das Gehirn als komplexer Computer zu verstehen ist, dessen Arbeitsweise im Prinzip geistige Prozesse insbesondere auch sprachlicher Art verständlich macht. Zweck der Nutzung ist hier nicht eigentlich das Resultat der Computerleistung, sondern Ablauf und Struktur dieser Leistung selbst, die die Prinzipien geistiger Prozesse erfaßbar macht.

(Manfred Bierwisch, Prinzipien einer kognitiven Orientierung der Computerlinguistik M.1988)

I.3. Montague–Grammatik (Geach: Hollywood semantics)

There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians; indeed, I consider it possible to comprehend the syntax and semantics of both kinds of languages within a single natural and mathematically precise theory.

(R.M.: Universal Grammar, in R.H.Thomason, ed., **Formal Philosophy**. Selected Papers of Richard Montague, New Haven 1974, 222-246 (222). of.p. (88))

I.4. Paradoxien und Puzzles in der analytischen Sprachphilosophie

(a) **Sprachakzeptizismus bei G.Frege**

Dies liegt also an einer Unvollkommenheit der Sprache, von der übrigens auch die Zeichensprache der Analysis nicht ganz frei ist; auch da können Zeichenverbindungen vorkommen, die den Anschein erwecken, als bedeuteten sie etwas, die aber wenigstens bisher noch bedeutungslos sind, z.B. divergente unendliche Reihen. (. . .) Von einer logisch vollkommenen Sprache (Begriffsschrift) ist zu verlangen, daßjeder Ausdruck, der aus schon eingeführten Zeichen in grammatisch richtiger Weise als Eigenname gebildet ist, auch in der Tat einen Gegenstand bezeichne, und daßkein Zeichen als Eigenname neu eingeführt werde, ohne daßihm eine Bedeutung gesichert sei. (. . .) Der demagogische Mißbrauch liegt hierbei ebenso nahe, vielleicht näher als bei vieldeutigen Wörtern. "Der Wille des Volkes" kann als Beispiel dazu dienen; dann daß es wenigstens keine allgemein angenommene Bedeutung dieses Ausdrucks gibt, wird leicht festzustellen sein.

(G.F.: Sinn und Bedeutung. Zs. für Philosophie u. philosophische Kritik, NF 100 (1892), 25-50 (41))

Man sieht heraus, we leicht man durch die Sprache zu falschen Auffassungen verleitet wird, und welchen Wert es daher für die Philosophie haben muss, sich der Herrschaft der Sprache zu entziehen. Wenn man versucht, auf ganz anderen Grundlagen mit ganz anderen Mitteln ein Zeichensystem zu errichten, wie ich es bei der Erfindung meiner Begriffsschrift versucht habe, wird man auf die falschen Analogien in der Sprache sozusagen mit der Nase gestossen.

(G.F.: Dialog mit Pünjer über Existenz. II. Freges Nachwort, in H.Hermes et.al., eds., Gottlob, Frege Nachgelassene Schriften, Hamburg 1969, 60–75 (74s))

(b) **Tarski zur Wahrheitsdefinition in natürlichen Sprachen**

Zunächst bin ich bemüht, diejenigen Schwierigkeiten zum Ausdruck zu bringen, welche bei Lösungsversuchen des in Frage stehenden Problems bezüglich der natürlichen Umgangssprache zum Vorschein kommen; insbesondere legt die Analyse der bekannten Antinomie vom Lügner den Schlußnahe, daßauf dem Boden der Umgangssprache (und in Bezug auf sie) nicht nur eine exakte Definition, sondern auch die konsequente Anwendung des Wahrheitsbegriffes schlechthin unmöglich sind.

(A.T.: Der Wahrheitsbegriff in den Sprachen der deduktiven Disziplinen, Anzeiger der Akademie der Wissenschaften in Wien. Mathematisch-naturwiss.Klasse 69(1932) 23-25.)

Tarskis Analyse des Lügners:

i Jede Aussage A hat einen Namen $\overline{A}$ in der Sprache L.

ii Für jede Formel $B(x)$ kann eine Aussage A konstruiert werden, die äquivalent zu $B(\overline{A})$ ist.

iii Die folgenden Axiome gelten für das Wahrheitsprädikat $T(x)$:

$$(TA) \quad T(\overline{A}) \Longleftrightarrow A$$

wobei A eine beliebige Aussage aus der Sprache L ist.

$$\begin{array}{rcll} A & \Longleftrightarrow & \neg T(\overline{A}) & (ii) \\ T(\overline{A}) & \Longleftrightarrow & A & (TA) \\ T(\overline{A}) & \Longleftrightarrow & \neg T(\overline{A}) & \end{array}$$

(c) **Quines slingshot**

(i) Bestimmte Kennzeichnungen
δA := die Zahl x, für die entweder gilt $x = 1$ und A oder $x = 0$ und nicht A

(ii) Referentiale Transparenz

$$\frac{\begin{array}{l}\text{Peter glaubt, daß } A(t) \\ t = t'\end{array}}{\text{Peter glaubt, daß } A(t')}$$

(iii) Logische Transparenz

$$\frac{\begin{array}{l}\text{Peter glaubt, dass } A \\ \vdash A \Longleftrightarrow B\end{array}}{\text{Peter glaubt, dass } B}$$

Peter glaubt, daß Cicero Catilina bloßgestellt hat.

Peter glaubt, daß $\delta A = 1$ (ii)

(A= Cicero hat Catilina bloßgestellt)

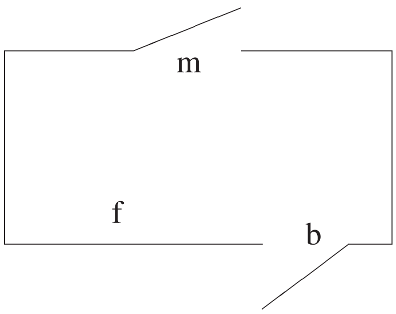
Peter glaubt, daß $\delta B = 1$

Peter glaubt, daß B

(Vgl. W.V.O. Quine, *Word and Object*, Cambridge, MA 1960)

(d) **Barwise und der Logiker K**

Mordfall Fred Smith (f)
Angeklagter: Brown(b)
Zeugin der Anklage: Mary Smith (m)
Sachverständiger der Verteidigung: K



Marys Zeugnis:

(i) m sah b durch Tür B eintreten.

(ii) f sah m durch die Tür F eintreten.

(iii) f sah b **nicht** durch Tür B eintreten.

Ks Argumentation:

$$\begin{array}{lll} & B(b) & \text{((i) \& Veridikalität visueller Perzeption)} \\ (*) & \neg(f \text{ sah } \neg B(b)) & \text{(Kontraposition)} \\ & f \text{ sah } (F(m) \text{ und } B(b)) \text{ oder } (F(m) \text{ und } \neg B(b)) & \\ & & ((ii) \& \vdash A \iff (A \text{ und } C) \text{ oder } (A \text{ und } \neg C)) \\ & f \text{ sah } F(m) \text{ und } \neg B(b) & \\ & & \text{(Widerspruch zu (*) wegen Veridikalität)} \\ & f \text{ sah } F(m) \text{ und } B(b) & \\ & & \text{(K weist Inkonsistenz in Marys Aussage nach!)} \end{array}$$

(Jon Barwise: Scenes and Other Situations, in J.B.: *The Situation in Logic*, Stanford: CLSI 1981)

(e) **Putnams Paradoxie**

(i) A cat is on a mat

(ii) A cat* is on a mat*

Mögliche Welten:
w: A und B
w': A und $\neg B$
w": $\neg A$ und B
A := Some cat is on some mat
B := Some cat is on some tree

	w	w'	w"
cat*	cherry	cat	cherry
mat*	tree	mat	quark

(Vgl. H.Putnam, *Reason, Truth and History*, Cambridge)

(f) **Hintikkas Any-These**

The word 'any' is acceptable (grammatical) in a given context X-any Y-Z if and only if an exchange of 'any' for 'every' results in a grammatical expression which is not identical in meaning with X-any Y-Z.

(J.Hintikkas: Quantifiers in Natural Languages, in E.Saarinen, ed., *Game-Theoretical Semantics*, Dordrecht 1979,81-118 (99))

(i) if any member contributes, I'll be happy

(ii) if every member contributes, I'll be happy

(iii) if Jane wins, anybody is happy

(iv) if Jane wins everybody is happy

(H) Grammatikalität abhängig von Bedeutungsgleichheit. Unentscheidbarkeit der semantischen Gültigkeit von Aussagen des Prädikatenkalküls erster Stufe.

(g) **Kripkes Wittgenstein**

Modern transformational linguistics, inasmuch as it explains all my specific utterances by my "grasp" of syntactic and semantic rules generating infinitely many sentences with their interpretation, seems to give an explanation of the type Wittgenstein would not permit. For the explanation is not in terms of my actual "performance" as a finite (and fallible) device. It is not a purely causal (neurophysiological) explanation . . . On the other hand, some aspects of Chomky"s views are very congenial to Wittgenstein"s conception. In particular, according to Chomsky, highly species-specific constraints - a "form of life" - lead a child to project, on the basis of exposure to a limited corpus of sentences, a variety of new sentences for new situations. There is no a priori inevitability in the child"s going on in the way he does, other than this is what the species does . . .
The point is that our understanding of the notion of competence is dependent on our understanding of the idea of "following a rule", . . . Wittgenstein would reject the idea that "competence" can be defined in terms of an idealized dispositional or mechanical model, and used without circularity to explicate the notion of following a rule.

For if statements attributing rule-following are neither to be regarded as stating facts, nor to be thought of as explaining our behavior, ... it would seem that the use of the ideas of rules and of competence in linguistics needs serious reconsideration, even if these notions are not rendered "meaningless".
These questions would arise even if, as throughout the present text, we deal with rules, like addition, that are stated explicitly. These rules we think of ourselves as grasping consciously; in the absence of Wittgenstein"s sceptical arguments, we would see no problem in the assumption that each particular answer we produce is justified by our "grasp" of the rules. The problems are compounded if, as in linguistics, the rules are thought of as tacit, to be reconstructed by the scientist and inferred as an explanation of behavior.

(S. Kripke, Wittgenstein on Rules and Private Language, Cambridge, MA 1982, p.97 note, p.31 note.)

I.5. Die Chomsky-Hierarchie

Generative Grammatiken werden aufgefaßt als formale Systeme, die

1. die **wohlgeformten** Ketten oder **Sätze** innerhalb der Menge aller Ketten über einem Vokabulat Σ definieren und die
2. den wohlgeformten Ketten eine oder mehrere **Strukturbeschreibungen** zuweisen.

Die Menge der wohlgeformten Ketten einer Grammatik G wird als ihre **schwache** generative Kapazität und die Menge der Strukturbeschreibungen als ihre **starke** generative Kapazität bezeichnet.

Generative Grammatik

Eine **generative Grammatik** ist ein Quadrupel $G = (\Sigma, F, S, P)$, wobei

- Σ eine endliche Menge von **Terminalsymbolen**,
- F eine endliche Menge an **Hilfssymbolen** $(F \cap \Sigma = \emptyset)$,
- P eine endliche Menge von **Regeln** $l \to r \quad (l, r \in (\Sigma \cup F)^*)$ und
- S das **Startsymbol** ist $(S \in F)$.

Die Menge $L(G) = \{w \in \Sigma^* \mid S \xrightarrow{*} w\}$ heißt die von G erzeugte **Sprache**.

Eine Grammatik $G = (\Sigma, F, S, P)$ ist

a) vom Typ 0, wenn P keine Regel der Form

$$\lambda \to r$$

enthält (λ das leere Wort).

b) vom Typ 1, wenn alle Regeln aus P die Form

$$uAv \to uwv$$

haben.

c) vom Typ 2, wenn alle Regeln aus P die Form

$$A \to r$$

haben.

d) vom Typ 3, wenn alle Regeln aus P die Form

$$A \to aB,\ A \to B,\ A \to a,\ A \to \lambda$$

oder die Form

$$A \to Ba,\ A \to B,\ A \to a,\ A \to \lambda$$

haben, wobei $A, B \in F$, $a \in \Sigma$ und $u, v, w, r \in (\Sigma \cup F)^*$ sind.

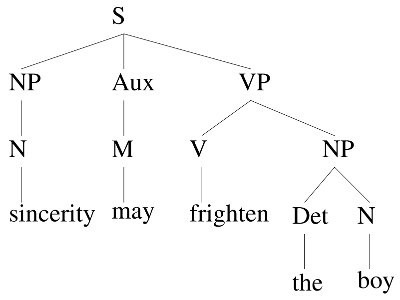
Chomsky-Hierarchie

Grammatiken vom Typ 0,1,2 und 3 werden auch als **Ersetzungssysteme**, **kontextsensitive**, **kontextfreie** bzw. **reguläre** Grammatiken bezeichnet.

$$\mathcal{L}_0 \not\supseteq \mathcal{L}_1, \not\supseteq \mathcal{L}_2, \not\supseteq \mathcal{L}_3$$

Ein Beispiel

S $\rightarrow$ NP Aux VP	Det $\rightarrow$ the
VP $\rightarrow$ V NP	M $\rightarrow$ may
NP $\rightarrow$ Det N	N $\rightarrow$ sincerity
NP $\rightarrow$ N	N $\rightarrow$ boy
Aux $\rightarrow$ M	V $\rightarrow$ frighten



Hypothese

Natürliche Sprachen sind eine **echte** Untermenge der kontextsensitiven Sprachen.

$$L_2 = \{a^n b^n\} \qquad \text{(Klammersprache)}$$
$$L_2 \in \mathcal{L}_2$$
$$S \rightarrow aSb \quad S \rightarrow ab$$

Schleifenlemma

Sei L eine reguläre Sprache über Σ. Dann gibt es eine Konstante k (abhängig von L), für die gilt, daß es für alle $w \in L$ mit $|w| \geq k$ Ketten $u, v, z \in \Sigma^*$ gibt mit folgenden Eigenschaften:

(i) $w = uvz$

(ii) $|av| \leq k$

(iii) $|v| \geq 1$

(iv) $uv^n z \in L$ für alle $n \geq 0$

Folgerung

L_2 ist nicht regulär.

Beispiele (Chomsky, Peters)

(i) She is more afraid than you are.

(ii) She is as afraid as you are.

(iii) She is as certain that she is more dismayed that cats bark than you are as you are.

(iv) Die Maus, die die Katze, die der Hund verfolgte, jagte, fraß den Käse.

Folgerung

Natürliche Sprachen sind nicht regulär.

Evidenz für Kontextsensitivität natürlicher Sprachen

Bambara	(ww)
Schweizerdeutsch	$(a^n b^m c^n d^m)$

Schleifenlemma (für kontextfreie Sprachen)
Für jede kontextfreie Sprache L gibt es Konstanten p und q, so daß es für jeden Satz $z \in L$ ($|z| \geq p$) Ketten $u, v, w, x, y \in \Sigma^*$ mit folgenden Eigenschaften gibt:

(i) $z = uvwxy$

(ii) $|vwx| \leq q$

(iii) $|vx| \geq 1$

(iv) $uv^n wx^n y \in L$ für alle $n \geq 0$

Folgerung
Natürliche Sprachen sind nicht kontextfrei.

Empirische Evidenz

Bisher sind keine grammatischen Phänomene entdeckt worden, die über den Bereich einer “milden” Kontextsensitivität hinausgehen.

Outline Part II

- Motivation and Theoretical Background
- Context-Free Tree Grammars
- LIFT-ing
- Coding of LIFT-ed CF Tree Languages
- Back to the Intended Structures
- Coding of the Intended Structures

Linguistic Background

J. Rogers (1994,98) gives a formalization within MSO tree logic of a main paradigm in Theoretical Linguistics:

The Principle and Parameters Approach

in the tradition of N. Chomsky

+ Natural Formalization of (almost) all Principles
+ Modular
+ Decidable
+ Descriptive Complexity Result

− Undecidability of *Free Indexation* (Grid)

− Non Context-Free Phenomena in Natural Language

Rogers (1994,1998): $L^2_{K,P}$

- K a set of individual constants
- P a set of predicate constants
- immediate dominance: $\triangleleft$
- reflexive dominance: $\triangleleft^*$
- proper dominance: $\triangleleft^+$
- proper precedence: $\prec$
- usual connectives, variables, equality, set membership and quantifiers:
  - a set of *monadic* second-order variables (X,Y,...);
  - a set of individual variables (x,y,...);
  - boolean connectives: $\neg$, $\wedge$, ...;
  - Quantifiers over individuals and sets: $\exists$, $\forall$;
  - equality and membership: $\approx$, $\in$.

Linguistic Motivation

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

German: Palindrome language—CF

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

Dutch: $a^n b^n$—CF

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

Swiss-German: $a^n b^m c^n d^m$—Non-CF

Formal Background

Weak Equivalence of Mildly Context-Sensitive Formalisms:

$\mathrm{STR(HR)}$: languages generated by string generating hyperedge replacement grammars

$\mathrm{OUT(DTWT)}$: output languages of deterministic tree-walking tree-to-string transducers

$\mathrm{yDT_{fc}(REGT)}$: yields of images of the regular tree languages under deterministic finite-copying top-down tree transductions

MCFL: languages generated by multiple context-free grammars

MCTAL: languages generated by multi-component tree adjoining grammars

LCFRL: languages generated by linear context-free rewriting systems

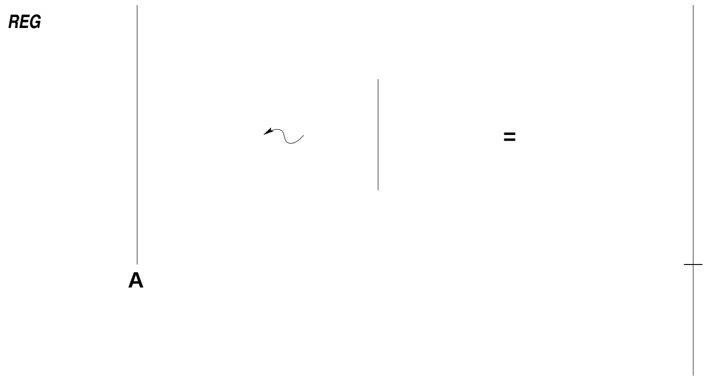
LUSCL: languages generated by local unordered scattered context grammars

General Structure of the Approach

LIFT-ing of Context-Free Tree Languages

$\Rightarrow$ Regular Characterization of the LIFT-ed Languages

$\Rightarrow$ Retrieval of the Intended Structures via Regular Relations



Definition (word models). Let Σ be an alphabet and let $\tau(\Sigma)$ be the vocabulary $\{<\} \cup \{P_a \mid a \in \Sigma\}$, where $<$ is binary and the P_a are monadic. A **word model** for $u \in \Sigma^*$ is a structure of the form

$$M = (B, <, P_a)$$

where $|B| = \mathrm{length}(u)$, $<$ is an ordering of B and the P_a correspond to positions in u carrying the label a:

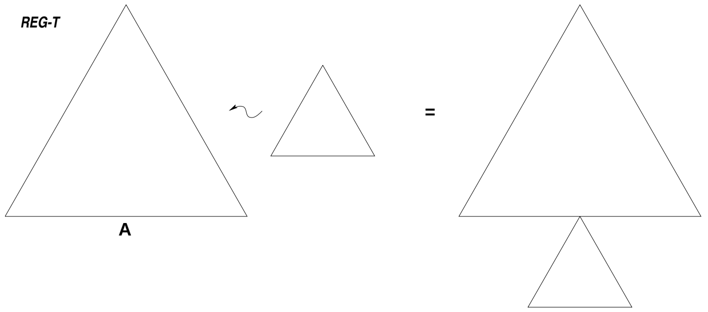
$$P_a = \{b \in B \mid \mathrm{label}(b) = a\}.$$

Theorem (Elgot). L is regular over the alphabet Σ iff L is definable in monadic second order logic over the vocabulary $\tau(\Sigma)$. More succinctly:

$L \in \mathcal{L}_{\mathsf{Reg}}$ iff

$\exists A \in MSO(\Sigma)$ such that

$$L = Mod(A) = \{u \in \Sigma^* \mid u \models A\}.$$



Definition (labeled tree models). Let Σ be a ranked alphabet and let $\tau(\Sigma)$ be the vocabulary $\{succ_1, \ldots, succ_n\} \cup \{P_a \mid a \in \Sigma\}$, where the $succ_i$ are binary and the P_a monadic. A **tree model** for $t \in T_\Sigma$ is a structure of the form

$$M = (B, succ_i, <, P_a)_{1 \leq i \leq n}$$

where B is in one-one correspondence with $dom(t)$, the tree domain of t, the $succ_i$ ($i = 1, \ldots, n$) are the successor functions over the tree domain, $<$ is the prefix order and the P_a correspond to positions with label a:

$$P_a = \{b \in B \mid label(b) = a\}.$$

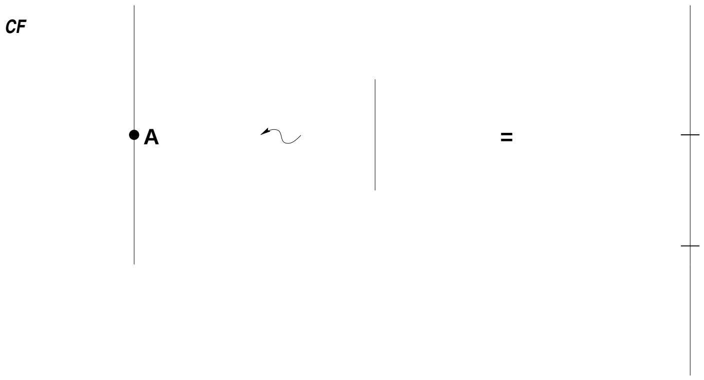
Theorem (Doner, Thatcher–Wright). Let Σ be a ranked alphabet. $L \subseteq \Sigma^*$ is **context-free** iff L is the yield of a set of tree models definable in monadic second order logic over the vocabulary $\tau(\Sigma)$.

$L \in \mathcal{L}_{CF}$ iff

$\exists A \in MSO(\Sigma)$ such that

$$L = yield(Mod(A))$$

where $Mod(A) = \{t \in T_\Sigma \mid t \models A\}$.

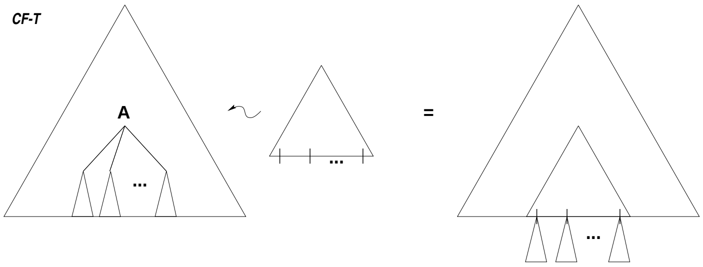


Theorem (Lautemann, Schwentick, Thérien). L is context-free over the alphabet Σ iff L is definable in dyadic second-order logic over the vocabulary $\tau(\Sigma)$ by a Σ_1^1-sentence where the quantified dyadic variables range over matchings. More succinctly:

$L \in \mathcal{L}_{\mathsf{CF}}$ iff

$$\exists A \in \Sigma_1^1 - \mathrm{DSO}(\Sigma) \text{ such that}$$

$$L = \mathrm{Mod}(A) = \{u \in \Sigma^* \mid u \models A\}.$$



Conjecture. Let Σ be a ranked alphabet. $L \subseteq T_\Sigma$ is a context-free tree language iff L is definable in dyadic second-order logic over the vocabulary $\tau(\Sigma)$ by a Σ_1^1-sentence where the quantified dyadic variables range over matchings. More succinctly:

$L \in \mathcal{L}_{\mathsf{CF\text{-}TL}}$ iff

$$\exists A \in \Sigma_1^1 - \mathrm{DSO}(\Sigma) \text{ such that}$$

$$L = \mathrm{Mod}(A) = \{t \in T_\Sigma \mid t \models A\}.$$

Tree Grammars

Definition (Tree Grammar). A *context-free tree grammar* Γ is a 5-tuple $\langle \Sigma, \mathbf{F}, S, \mathbf{X}, \mathbf{P} \rangle$, where

$\Sigma, \mathbf{F}$	are the ranked alphabets of *inoperatives* and *operatives*;
$S \in \mathbf{F}$	is the starting symbol;
$\mathbf{X}$	is a countable set of variables; and
$\mathbf{P}$	is a set of productions.

The $p \in \mathbf{P}$ have the form

$$F(x_i, \cdots, x_n) \longrightarrow t$$

with $F \in \mathbf{F}$, $x_i, \cdots, x_n \in \mathbf{X}$, and t a term over $\Sigma \cup \mathbf{F} \cup \{x_i, \cdots, x_n\}$.

An application of a rule $F(x_i, \cdots, x_n) \longrightarrow t$ "rewrites" a tree rooted in F as the tree t with its respective variables substituted by F's daughters.

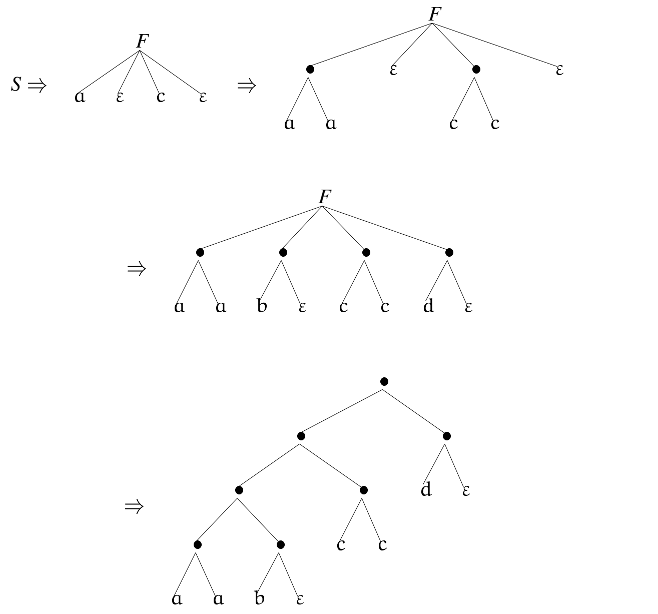
Tree grammars with $\mathbf{F}_n = \emptyset, n \neq 0$, are called *regular*.

A CF-Tree Grammar for $a^n b^m c^n d^m$: $\mathcal{G}$

$$\Sigma_0 = \{\varepsilon, a, b, c, d\}$$
$$\Sigma_2 = \{\bullet\}$$
$$\mathbf{F}_0 = \{S\} \qquad \mathbf{F}_4 = \{F\}$$
$$\mathsf{S} = S \qquad \mathbf{X} = \{x_1, x_2, x_3, x_4\}$$

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

An example derivation of $aabccd$



Definition (LIFT). Suppose that Σ is a ranked alphabet. The *derived* $\mathbb{N}$-sorted alphabet Σ^L is defined as follows: For each $n \geq 0$, $\Sigma'_n = \{f' \mid f \in \Sigma_n\}$ is a new set of symbols of sort n; for each $n \geq 1$ and each i, $1 \leq i \leq n$, π_i^n is a new symbol, the ith projection symbol of sort n; for each $n \geq 0, k \geq 0$ the new symbol $c_{n,k}$ is the (n,k)th composition symbol. The set of all $c_{n,k}$ will be denoted by $\mathbf{C}$.

$$\Sigma_0^L = \Sigma'_0$$
$$\Sigma_n^L = \Sigma'_n \cup \{\pi_i^n \mid 1 \leq i \leq n\} \text{ for } n \geq 1$$
$$\Sigma_{n,k}^L = \{c_{n,k}\} \text{ for } n, k \geq 0$$
$$\Sigma_i^L = \emptyset \text{ otherwise}$$

For $k \geq 0$, $\text{LIFT}_k^\Sigma : T(\Sigma, \mathbf{X}_k) \to T(\Sigma^L, k)$ is defined as follows:

$$\text{LIFT}_k^\Sigma(x_i) = \pi_i^k$$
$$\text{LIFT}_k^\Sigma(f) = c_{0,k}(f') \text{ for } f \in \Sigma_0$$
$$\text{LIFT}_k^\Sigma(f(t_1, \ldots, t_n)) =$$
$$c_{n,k}(f', \text{LIFT}_k^\Sigma(t_1), \ldots, \text{LIFT}_k^\Sigma(t_n))$$
$$\text{for } f \in \Sigma_n, n \geq 1$$

A LIFT-ed Regular Tree Grammar: $\mathcal{G}'$

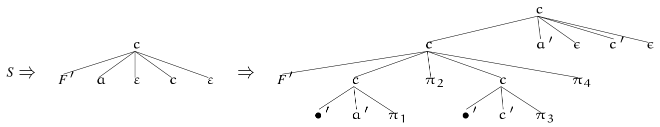
$$\Sigma_0^L = \{\varepsilon, a', b', c', d', S'\} \qquad \Sigma_2^L = \{\bullet'\}$$
$$\Sigma_4^L = \{F'\} \cup \{\pi_1, \pi_2, \pi_3, \pi_4\} \qquad \Sigma_{n,k}^L = \{c\} \text{ for simplicity}$$
$$\mathbf{F}_0 = \{S', F'\} \qquad \mathsf{S} = S' \qquad \mathbf{X} = \emptyset$$

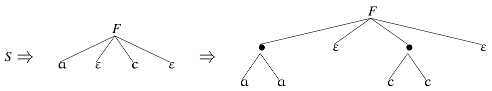
$$\mathbf{P} = \left\{ \begin{array}{lcl} S' & \longrightarrow & \varepsilon \\ S' & \longrightarrow & c(F', a', \varepsilon, c', \varepsilon) \\ S' & \longrightarrow & c(F', \varepsilon, b', \varepsilon, d') \\ F' & \longrightarrow & c(F', c(\bullet', a', \pi_1), \pi_2, c(\bullet', c', \pi_3), \pi_4) \\ F' & \longrightarrow & c(F', \pi_1, c(\bullet', b', \pi_2), \pi_3, c(\bullet', d', \pi_4)) \\ F' & \longrightarrow & c(\bullet', c(\bullet', c(\bullet', \pi_1, \pi_2), \pi_3), \pi_4) \end{array} \right\}$$

$$\Sigma_0 = \{\varepsilon, a, b, c, d\} \quad \Sigma_2 = \{\bullet\} \quad \mathbf{F}_0 = \{S\} \quad \mathbf{F}_4 = \{F\} \quad \mathsf{S} = S \quad \mathbf{X} = \{x_1, x_2, x_3, x_4\}$$

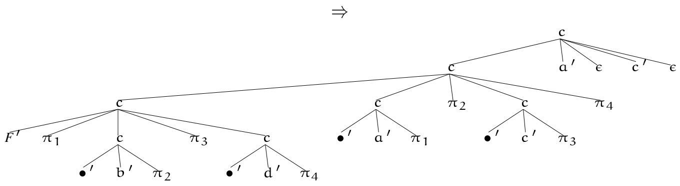
$$\mathbf{P} = \left\{ \begin{array}{rcl} S & \longrightarrow & \varepsilon \\ S & \longrightarrow & F(a, \varepsilon, c, \varepsilon) \\ S & \longrightarrow & F(\varepsilon, b, \varepsilon, d) \\ F(x_1, x_2, x_3, x_4) & \longrightarrow & F(\bullet(a, x_1), x_2, \bullet(c, x_3), x_4) \\ F(x_1, x_2, x_3,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

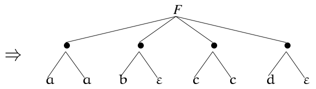
A sample derivation I



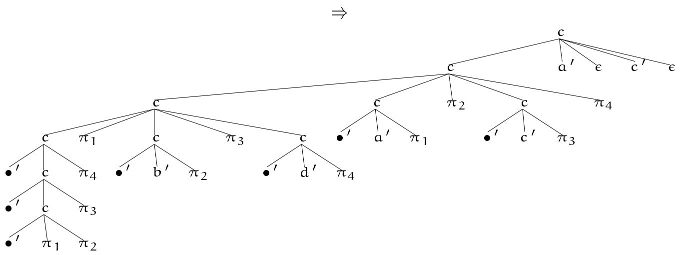


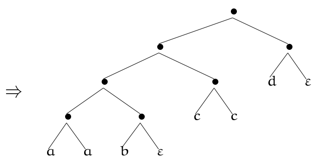
A sample derivation II





A sample derivation III





Definition (Bottom-Up Tree Automaton).

$$\mathfrak{A} = \langle A, \Sigma, \delta, a_0, F \rangle$$

- A the (finite) set of states,
- Σ a ranked alphabet,
- $a_0 \in A$ the initial state,
- $F \subseteq A$ the final states and
- $\delta : (A^n \times \Sigma_n) \to A$ the transition function.

$$\begin{aligned} h_\delta(\varepsilon) &= a_0 \\ h_\delta(\sigma(t_1, \ldots, t_n)) &= \delta(h_\delta(t_1), \ldots, h_\delta(t_n), \sigma) \\ & \qquad t_i \in T_\Sigma, 1 \leq i \leq n, \sigma \in \Sigma_n \end{aligned}$$

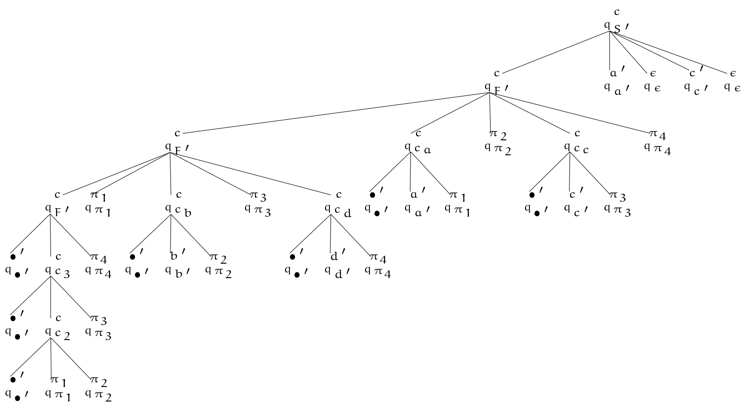
$$T(\mathfrak{A}) = \{t \mid h_\delta(t) \in F\}$$

A Tree Automaton for the LIFT-ed CF Tree Grammar

$$\mathfrak{A}_{\mathcal{G}'} = \langle Q, \Sigma, \delta, q_0, \{q_{S'}\} \rangle$$

$$\begin{array}{lll@{\qquad}lll}
(q_0,q_0,q_0,q_0,q_0,\epsilon) & \to & q_{\epsilon} & (q_0,q_0,q_{\bullet'},q_{a'},q_{\pi_1},c) & \to & q_{c_a} \\
(q_0,q_0,q_0,q_0,q_0,a') & \to & q_{a'} & (q_0,q_0,q_{\bullet'},q_{c'},q_{\pi_3},c) & \to & q_{c_c} \\
(q_0,q_0,q_0,q_0,q_0,b') & \to & q_{b'} & (q_0,q_0,q_{\bullet'},q_{b'},q_{\pi_2},c) & \to & q_{c_b} \\
(q_0,q_0,q_0,q_0,q_0,c') & \to & q_{c'} & (q_0,q_0,q_{\bullet'},q_{d'},q_{\pi_4},c) & \to & q_{c_d} \\
(q_0,q_0,q_0,q_0,q_0,d') & \to & q_{d'} & (q_0,q_0,q_{\bullet'},q_{\pi_1},q_{\pi_2},c) & \to & q_{c_2} \\
(q_0,q_0,q_0,q_0,q_0,\pi_1) & \to & q_{\pi_1} & (q_0,q_0,q_{\bullet'},q_{c_2},q_{\pi_3},c) & \to & q_{c_3} \\
(q_0,q_0,q_0,q_0,q_0,\pi_2) & \to & q_{\pi_2} & (q_0,q_0,q_{\bullet'},q_{c_3},q_{\pi_4},c) & \to & q_{F'} \\
(q_0,q_0,q_0,q_0,q_0,\pi_3) & \to & q_{\pi_3} & (q_{F'},q_{\pi_1},q_{c_b},q_{\pi_3},q_{c_d},c) & \to & q_{F'} \\
(q_0,q_0,q_0,q_0,q_0,\pi_4) & \to & q_{\pi_4} & (q_{F'},q_{c_a},q_{\pi_2},q_{c_c},q_{\pi_4},c) & \to & q_{F'} \\
(q_0,q_0,q_0,q_0,q_0,\bullet') & \to & q_{\bullet'} & (q_{F'},q_{\epsilon},q_{b},q_{\epsilon},q_{d},c) & \to & q_{S'} \\
(q_0,q_0,q_0,q_0,q_0,\epsilon) & \to & q_{S'} & (q_{F'},q_{a},q_{\epsilon},q_{c},q_{\epsilon},c) & \to & q_{S'}
\end{array}$$

An Example Run



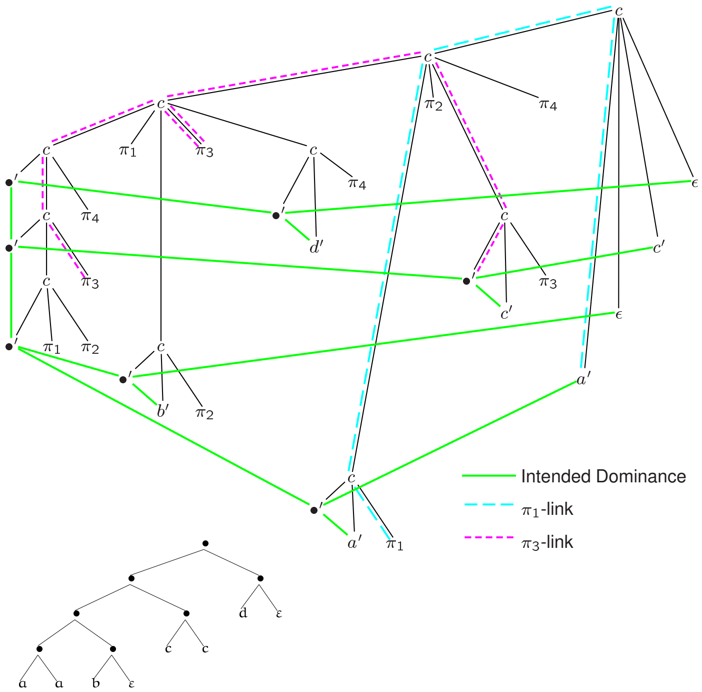
Goal:

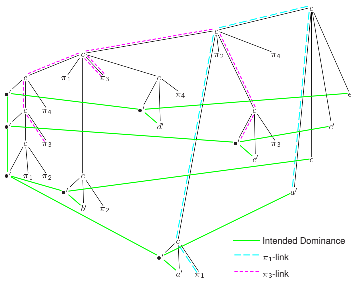
Define a set of relations
$R^I = \{\blacktriangleleft^\star$ *(dominance)*, $\blacktriangleleft$, $\blacktriangleleft^+$, $\lhd$ *(precedence)*, *c-command*, $\ldots\}$
holding between the nodes $n \in N^L$ of the explicit tree T^L
which carry a *"linguistic"* label $l \in \mathcal{T}^L \cup \mathcal{N}^L$ ($\subseteq \bigcup_{n\geq 0} \Sigma_n$)
in such a way, that when interpreting

$\blacktriangleleft^\star \in R^I$ as a tree order on the set of *"linguistic"* nodes and
$\lhd \in R^I$ as the precedence relation on the resulting structure,

$\langle\{n \mid n \in N^L \wedge \ell(n) \in \mathcal{T}^L \cup \mathcal{N}^L\}, \blacktriangleleft^\star, \lhd\rangle$
is in fact the intended tree corresponding to T^L.

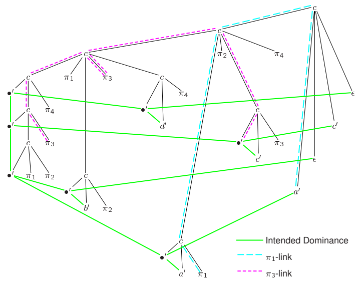
Reconstructing the Intended Structures





A few general facts

1. Our trees feature three families of labels: the "linguistic" symbols $\mathbf{L} = \mathcal{T}^L \cup \mathcal{N}^L$, where $\mathcal{T} = \Sigma_0$ of the underlying macro-grammar and $\mathcal{N} = \bigcup_{n \geq 1} \Sigma_n$; the "composition" symbols $\mathbf{C} = \bigcup_{n,k \geq 0} C_{n,k}$; and the "projection" symbols Π.
2. All non-terminal nodes in T^L are labeled by some $c \in \mathbf{C}$. No terminal node is labeled by some c.
3. The terminal nodes in T^L are either labeled by some "linguistic" symbol or by some "projection" symbol $\pi \in \Pi_i$.



A few facts (cont'd.)

4. Any "linguistic" node dominating anything in the intended tree is on some left branch in T^L, i.e., it is the left-most daughter^L of some $c \in \mathbf{C}$.
5. For any node p labeled with some "projection" symbol $\pi \in \Pi_i$ in T^L there is a unique node n (labeled with some $c \in \mathbf{C}$ by (2.)) which properly dominates p and whose i-th sister will eventually evaluate to the value of π. Moreover, n will be the first node properly dominating p which is on a left branch.

MSO definable transductions: $\mathcal{R} \rightsquigarrow \mathcal{Q}$

$$(\varphi, \psi, (\theta_q)_{q \in Q})$$

φ the domain of the transduction

ψ the resulting domain of $\mathcal{Q}$

θ_q the new relations

Definition (Tree-Walking Automaton). A *tree-walking automaton (with tests)* over some ranked alphabet Σ is a finite automaton $\mathfrak{A} = (Q, \Delta, \delta, I, F)$ with states Q, directives Δ, transitions $\delta : Q \times \Delta \to Q$ 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)",
- $\varphi(x)$ "verify that φ holds at the current node".

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

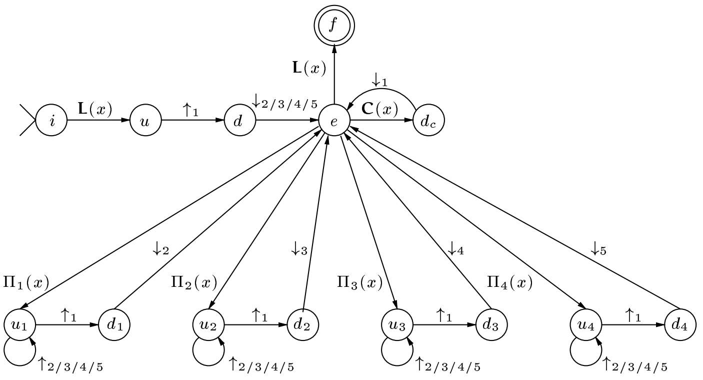
$$R_t(\mathfrak{A}) = \{(x, y) \mid (x, q_i) \stackrel{*}{\Longrightarrow} (y, q_f) \text{ for some } q_i \in I \text{ and some } q_f \in F\}$$

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 $\varphi(x)$ of $\mathfrak{A}$ are MSO definable, $\mathfrak{A}$ specifies a *regular tree-node relation*, which is itself MSO definable (B & E 1997).

The tree-walking automaton $A_{\blacktriangleleft}$



The walking language $W_{\blacktriangleleft}$

$$
\begin{aligned}
W_{\blacktriangleleft} = {} & \mathbf{L}(x) \cdot \uparrow_1 \cdot (\downarrow_2 \cup \downarrow_3 \cup \downarrow_4 \cup \downarrow_5) \cdot \\
& (W_{\mathbf{C}} \cup W_{\Pi_1} \cup W_{\Pi_2} \cup W_{\Pi_3} \cup W_{\Pi_4})^* \cdot \mathbf{L}(x)
\end{aligned}
$$

$$
\begin{aligned}
W_{\mathbf{C}} &= \mathbf{C}(x) \cdot \downarrow_1 \\
W_{\Pi_1} &= \Pi_1(x) \ \cdot \ (\uparrow_2 \cup \uparrow_3 \cup \uparrow_4 \cup \uparrow_5)^* \cdot \uparrow_1 \cdot \downarrow_2 \\
W_{\Pi_2} &= \Pi_2(x) \ \cdot \ (\uparrow_2 \cup \uparrow_3 \cup \uparrow_4 \cup \uparrow_5)^* \cdot \uparrow_1 \cdot \downarrow_3 \\
W_{\Pi_3} &= \Pi_3(x) \ \cdot \ (\uparrow_2 \cup \uparrow_3 \cup \uparrow_4 \cup \uparrow_5)^* \cdot \uparrow_1 \cdot \downarrow_4 \\
W_{\Pi_4} &= \Pi_4(x) \ \cdot \ (\uparrow_2 \cup \uparrow_3 \cup \uparrow_4 \cup \uparrow_5)^* \cdot \uparrow_1 \cdot \downarrow_5
\end{aligned}
$$

$W_{\blacktriangleleft}$ can be inductively translated into a MSO-formula $\mathit{trans}_{W_{\blacktriangleleft}}(x, y)$.

Reflexive transitive closure in MSO (on finite structures)

$$
\begin{aligned}
R\text{-}\mathit{closed}(X) &\equiv (\forall x, y)[x \in X \wedge R(x, y) \ \rightarrow \ y \in X] \\
[R^*(x, y) &\equiv (\forall X)[R\text{-}\mathit{closed}(X) \wedge x \in X \ \rightarrow \ y \in X]
\end{aligned}
$$

Walking language $\Rightarrow$ MSO-formula

$$
\begin{aligned}
\mathit{trans}_{\emptyset}(x, y) &= \bot \\
\mathit{trans}_{\downarrow_i}(x, y) &= \mathit{edge}_i(x, y) \\
\mathit{trans}_{\uparrow_i}(x, y) &= \mathit{edge}_i(y, x) \\
\mathit{trans}_{\varphi(x)}(x, y) &= \varphi(x) \wedge x = y \\
\mathit{trans}_{W_1 \cup W_2}(x, y) &= \mathit{trans}_{W_1}(x, y) \vee \mathit{trans}_{W_2}(x, y) \\
\mathit{trans}_{W_1 \cdot W_2}(x, y) &= \exists z \, (\mathit{trans}_{W_1}(x, z) \wedge \mathit{trans}_{W_2}(z, y)) \\
\mathit{trans}_{W^*}(x, y) &= \mathit{trans}^*_W(x, y) \\
\mathit{trans}^*_W(x, y) &= \forall X \, (\forall v, w \, (v \in X \wedge \mathit{trans}_W(v, w) \ \rightarrow \ w \in X) \\
&\qquad \wedge x \in X \ \rightarrow \ y \in X)
\end{aligned}
$$

Intended Structures via MSO transduction

$$(\varphi, \psi, (\theta_q)_{q \in Q})$$

$$Q = \{\blacktriangleleft, \blacktriangleleft^{\star}, \blacktriangleleft^{+}, \vartriangleleft, \ldots\}$$

$$\begin{aligned}
\varphi &= \text{ the formula coding } \mathfrak{A}_{\mathcal{G}'} \\
\psi &= (\exists y)[x \blacktriangleleft y \lor y \blacktriangleleft x] \\
\theta_{\blacktriangleleft} &= \text{ the formula for the walking language of } \mathfrak{A}_{\blacktriangleleft} \\
\theta_{\blacktriangleleft^{\star}} &= (\forall X)[\blacktriangleleft\text{-}closed(X) \lor x \in X \ \rightarrow \ y \in X] \\
\theta_{\blacktriangleleft^{+}} &= x \blacktriangleleft^{\star} y \lor x \not\approx y \\
\theta_{\vartriangleleft} &= \text{ another tree-walking automaton} \\
\theta_{\text{labels}} &= \text{ as in the old structure } R
\end{aligned}$$

Outlook

Theorem (Courcelle). The image of an MSO-definable class of structures under a definable transduction is not MSO definable in general.

But which class of structures is captured using regular means/MSO logic combined with regular/MSO definable binary node relations?

Is there a machine model for the MSO transduction such that we can implement the process?