Elements of Formal Linguistics for Computability Theorists

Tools, Concepts, Implications

Geoffrey K. Pullum

University of Edinburgh

June 2011

CiE, Sofia, Bulgaria

Introduction: human language

Humans can not only **entertain propositional thoughts** and **store, manipulate, and reflect** on them, but also

- **transmit** propositions overtly and intentionally
- with the **intent of altering epistemic states** of conspecifics
- **independently of stimulus control**

- **receive** them from conspecifics
- **acquire** new propositional information in that way

- using a **huge and apparently open** variety of signs with
- **arbitrary internal complexity**, and
- **learn** this system of signs rapidly and early

Introduction: human language

Other animals communicate, over various channels:

- **vision** (bodily shape, position, or color)
- **sound** (growling, squealing, calling)
- **odour**
- **touch**

and so on.

But they do not have **grammar**:

- phonology (sound patterns)
- morphology (word structure)
- semantics (literal meaning)
- syntax (expression structure)

⟵ MY FOCUS HERE

Grammaticality and ungrammaticality

Traditional grammars of English are massively **incomplete** and **imperfect**.

They will tell you that adjectives modify nouns, but they will never note that this is possible:

There are some ugly, ugly people out there.

I want you to be very, very, very careful.

We have many, many topics to discuss.

They will tell you that adverbs modify words other than nouns, but they will never warn you about this:

I eagerly opened the box.

**I opened eagerly the box.*

‘ * ’ indicates ‘not syntactically permissible (in Standard English)’

Grammaticality and ungrammaticality

They never point out that adverbs can be built up into phrases:

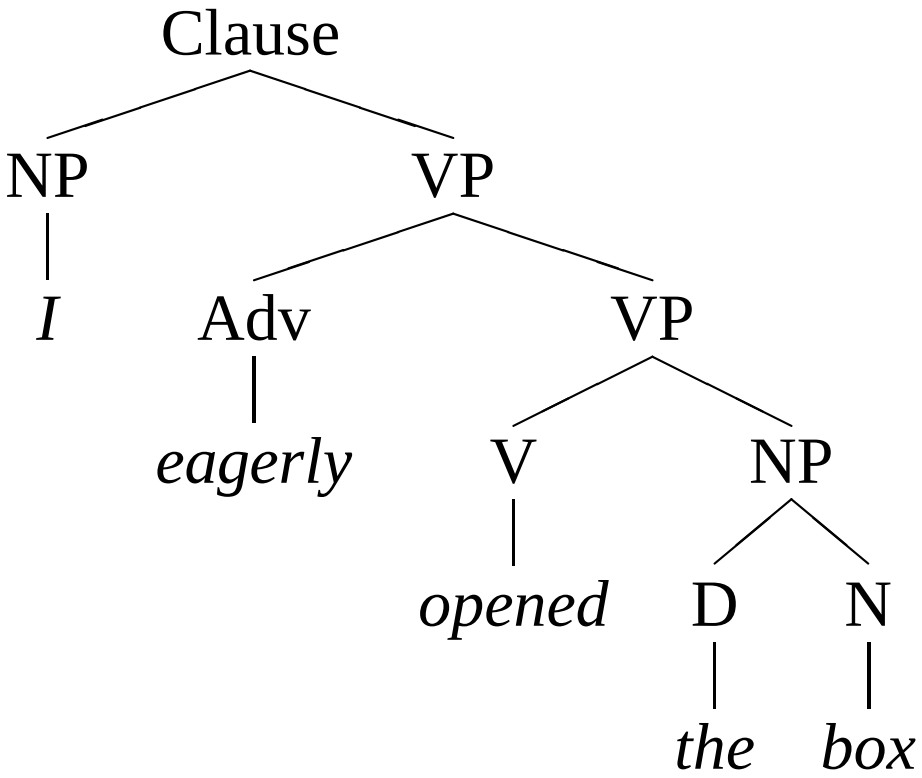
separately

[*separately* [*from* [*the rest*]]]

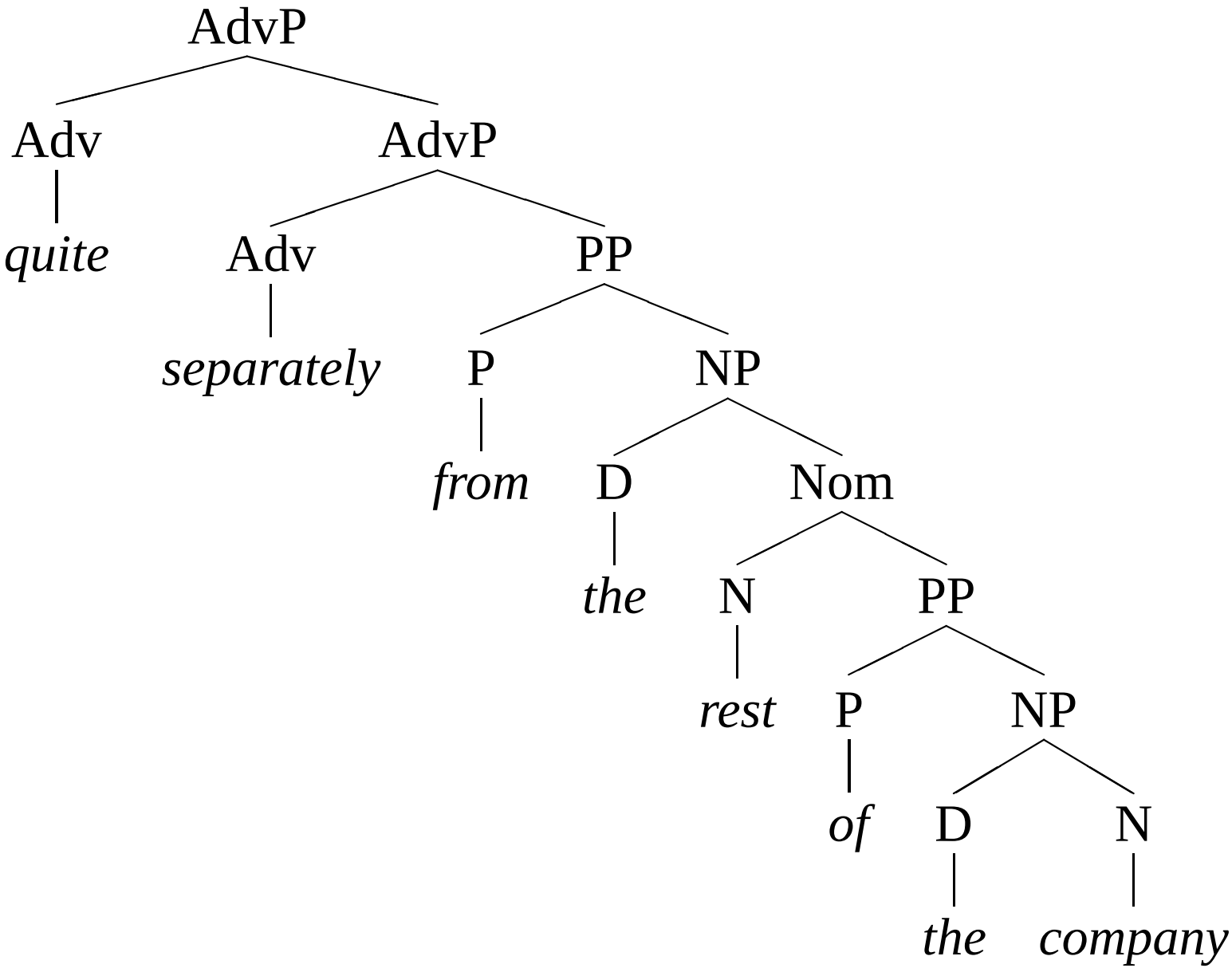
[*separately* [*from* [*the rest* [*of the company*]]]]

[*quite* [*separately* [*from* [*the rest* [*of the company*]]]]]

Tree diagrams of expression structure



Tree diagrams of expression structure



Generative grammars

Since 1957, linguists have become increasingly interested in **generative** grammars. These are grammars that (unlike traditional ones)

- presuppose nothing
- are fully explicit
- cover all details, both universal and parochial
- support semantics and phonology
- describe all and only properly formed expressions
- **are formulated as recursive specifications of set membership**

Origins of generative grammars

Emill Post's schema for canonical productions (1943):

$$g_{1_1} P_{i'_1} g_{1_2} P_{i'_2} \cdots g_{1_{m_1}} P_{i'_{m_1}} g_{1_{(m_1+1)}}$$
$$g_{2_1} P_{i''_1} g_{2_2} P_{i''_2} \cdots g_{2_{m_2}} P_{i''_{m_2}} g_{2_{(m_2+1)}}$$
$$\cdots\cdots\cdots\cdots$$
$$g_{k_1} P_{i_1^{(k)}} g_{k_2} P_{i_2^{(k)}} \cdots g_{k_{m_k}} P_{i_{m_k}^{(k)}} g_{k_{(m_k+1)}}$$

produce

$$g_1 P_{i_1} g_2 P_{i_2} \cdots g_m P_{i_m} g_{m+1}$$

- g_i are fixed strings given in the production
- P_i are free string variables
- strings matching lines 1 to k are sufficient to license the derivation of a string matching the last line

Canonical systems and computable enumerability

Post recognized that his canonical production systems provided specifications by enumeration for all computably enumerable (CE) sets.

Two theorems of Post's:

P1: All CE sets are still obtainable if productions are limited to

$g_1\ P$ produces $P\ g_2$

P2: All CE sets are still obtainable if productions are limited to

$P_1\ g_1\ P_2$ produces $P_1\ g_2\ P_2$

Below CE: the Chomsky Hierarchy

Assume finite vocabulary $V = V_N \cup V_T$, and rules ($\approx$ productions) of this form:

$$W\alpha Z \rightarrow W\beta Z \qquad (W, Z \in V^* \text{ and } \alpha \in V^* V_N V^*)$$

Now consider these restrictions on rule form:

Restriction 1: $\beta \neq \epsilon$

Restriction 2: $W = Z = \epsilon$

Restriction 3: $\beta \in V_T \vee \beta \in V_T V_N$

Below CE: the Chomsky Hierarchy

	Restriction 1	Restriction 2	Restriction 3
Type 0:	–	–	–
Type 1:	✓	–	–
Type 2:	✓	✓	–
Type 3:	✓	✓	✓

Grammar types: Type 0 $\supsetneq$ Type 1 $\supsetneq$ Type 2 $\supsetneq$ Type 3

Let $L(\text{Type } i) = \{L | \text{ some grammar of Type } i \text{ generates } L\}$

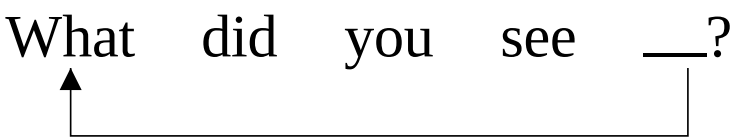
Hierarchy theorem (Chomsky 1959):

$$L(\text{Type 0}) = \text{Computably Enumerable (CE)} \supsetneq$$
$$L(\text{Type 1}) = \text{Context-Sensitive (CS)} \supsetneq$$
$$L(\text{Type 2}) = \text{Context-Free (CF)} \supsetneq$$
$$L(\text{Type 3}) = \text{Finite-State (FS)} = \text{Regular}$$

Automaton characterizations

TYPE	STRINGSETS	MACHINE TYPE
Type 0	CE	Turing machine
Type 1	CS	Linear Bounded Automaton
Type 2	CF	Pushdown Stack Automaton
Type 3	FS	Finite automaton

Transformations



T_{w_1}: Structural analysis: $X - NP - Y$ (X or Y may be null)
Structural change: $X_1 - X_2 - X_3 \rightarrow X_2 - X_1 - X_3$

In Post's notation:

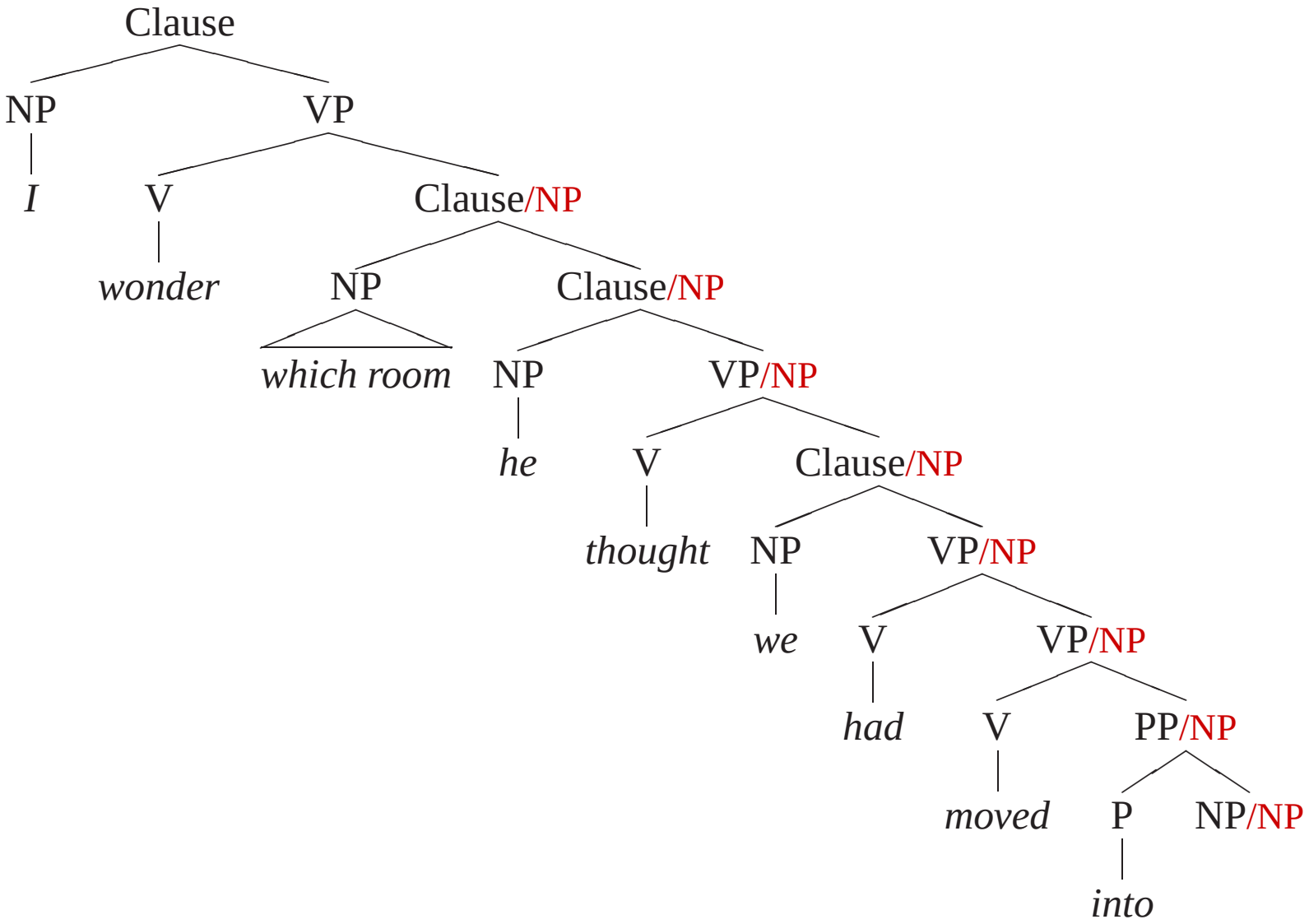
$P_1\ NP\ P_2$ produces $NP\ P_1\ P_2$

A remark by Hilary Putnam:

> Chomsky's general characterization of a transformational grammar is much too wide. It is easy to show that any recursively enumerable set of sentences could be generated by a transformational grammar in Chomsky's sense.

Transformations

But transformations may not be needed (Gazdar 1981):



What lies inside FS?

Type 0 (CE) $\supseteq$

Type 1 (CS) $\supseteq$

Type 2 (CF) $\supseteq$

Type 3 (FS) $\supseteq$

- - - - - - - - - - - - - - - - -

? ? ? ⟵ WHAT CLASSES LIE IN HERE?

- - - - - - - - - - - - - - - - -

Finite stringsets

Inside FS: the subregular classes

Finite State (FS)

- - - - - - - - - - - - - - - - -

Star Free (SF)

Locally Testable (LT)

Strictly Local (SL)

- - - - - - - - - - - - - - - - -

Finite stringsets

Strictly Local (SL) stringsets

Checking grammaticality in an SL_k stringset:

$$\underbrace{\sigma_1 \ \ldots \ (\ldots \text{Verb} \, ? \ldots) \ \ldots \ \sigma_{i_0}}_{\text{region now forgotten}} \ \overbrace{\sigma_{i_1} \ \ldots \ \sigma_{i_k}}^{\text{current scan}} \ \underbrace{\sigma_{i_k+1} \ \ldots \ (\ldots \text{Verb} \, ? \ldots) \ \ldots \ \sigma_n}_{\text{region yet to be checked}}$$

Locally Testable (LT) and Star-Free (SF)

The Locally Testable (LT) stringsets are the closure of SL under boolean operations.

This makes it possible to say that a string must contain exactly one occurrence of b.

The stringset a^*ba^* is not SL but it is LT.

The Star-Free (SF) stringsets are the closure of SL under boolean operations and concatenation.

This makes it possible to say that a string must contain exactly one occurrence of b and one of c, in that order.

The stringset $a^*ba^*ca^*$ is not LT but it is SF.

Star-Free (SF) stringsets

Definitions:

Asteration of a stringset L over V: $L^* =_{\text{def}} \bigcup_i L^i$.

Complement of a stringset L over V: $\overline{L} =_{\text{def}} V^* - L$.

E.g.: $a^*ba^*ca^*$ is SF, because it is denoted by this expression:

$$\overline{(b+c)} \cdot b \cdot \overline{(b+c)} \cdot c \cdot \overline{(b+c)}$$

'anything with no b or c, followed by b, followed by anything with no b or c, followed by c, followed by anything with no b or c'

What lies above CF?

Type 0 (CE) $\supseteq$

Type 1 (CS) $\supseteq$

- - - - - - - - - - - - - - - - -

? ? ? ⟵ WHAT CLASSES LIE IN HERE?

- - - - - - - - - - - - - - - - -

Type 2 (CF) $\supseteq$

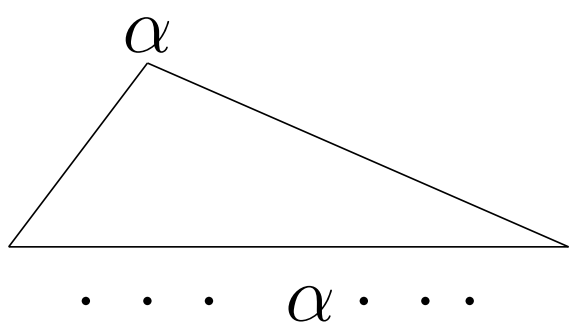
Type 3 (FS) $\supseteq$

Subregular classes

Finite stringsets

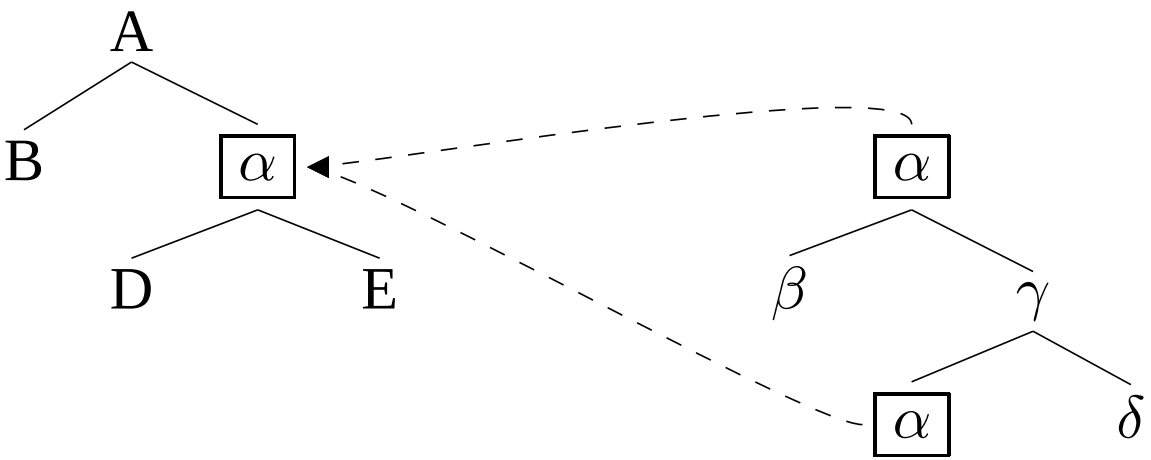
Tree Adjoining (TA) stringsets

An auxiliary tree:



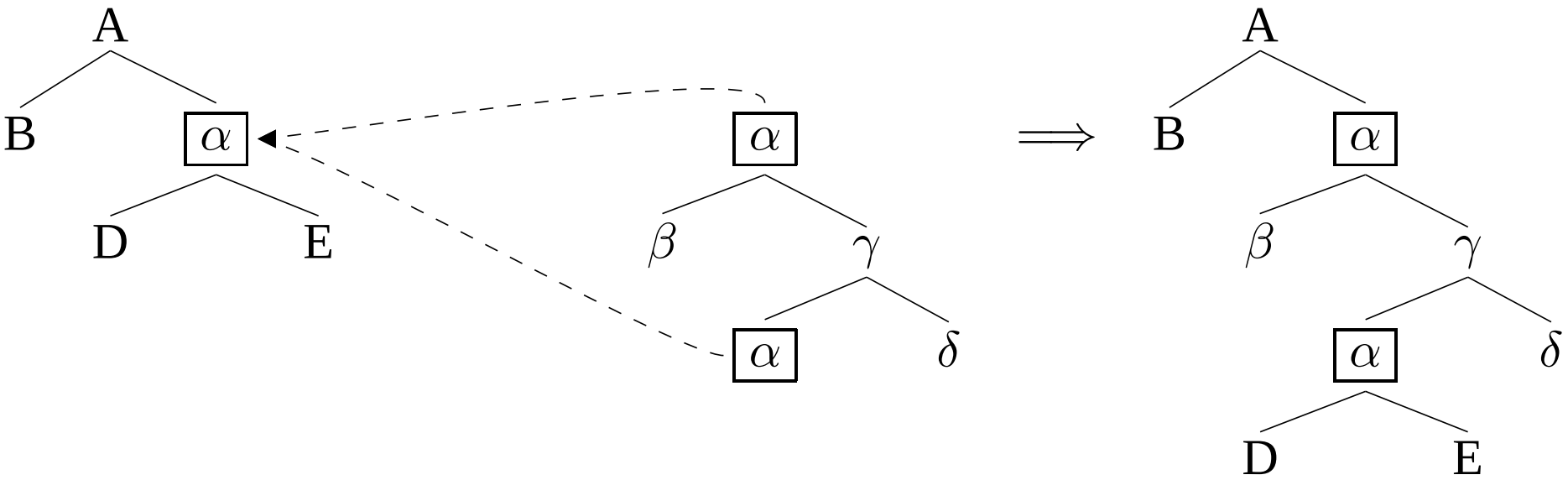
Tree Adjoining (TA) stringsets

Adjoining an auxiliary tree into another tree:



Tree Adjoining (TA) stringsets

Adjoining an auxiliary tree into another tree:



The Weir/Vijayshanker Theorem

Tree Adjoining stringsets (Joshi)

= Head-grammar stringsets (Roach)

= Combinatory Categorial stringsets (Steedman)

= Linear Indexed stringsets (Gazdar)

= Embedded PDA-recognizable (Vijayshanker)

Between FS and CE

Decidable

Primitive Recursive

Type 1 = Context-Free (CS)

$\cdots$ (multiple CF; growing CS; indexed) $\cdots$

Tree Adjoining (TA)

Type 2 = Context-Free (CF)

Deterministic Context-Free (D-CF)

Linear (LN)

Type 3 = Finite-State (FS)

$\cdots$ (subregular classes) $\cdots$

A more elaborate hierarchy

$$\text{CE} \supset \text{Decidable} \supset \text{Primitive Recursive} \supset$$

$$\text{CS} \supset \text{Indexed} \supset \text{Tree Adjoining} \supset$$

$$\text{Tree Adjoining} \supset \text{CF} \supset \text{Deterministic-CF} \supset$$

$$\text{Linear} \supset \text{FS} \supset \text{SF} \supset \text{LT} \supset \text{SL}$$

Mathematical questions of potential linguistic interest

- GENERATIVE CAPACITY of various forms of grammars (e.g., Can a Type *i* grammar generate any stringsets that cannot be generated by a grammar of Type *j*?)
- DECIDABILITY QUESTIONS for grammars of particular types (e.g., Is it decidable whether an arbitrary Type *i* grammar is ambiguous, or generates V^*, or generates anything at all?)
- the RECOGNITION PROBLEM (i.e., Is it decidable for an arbitrary grammar *G* and a string *w* whether *G* generates *w*?)
- 'LEARNABILITY' problems (e.g., Is there an algorithm that, given a stream of strings belonging to some stringset in a given class, will after a finite number of guesses correctly identify a grammar for it?)

Providing a grammar for every decidable stringset

Janssen, Kok & Meertens (1977)

- **Theorem 1** There is **no** CE set of generative grammars containing a grammar for every decidable stringset and associated with a procedure for deciding membership.
- **Theorem 2** There is a CE (in fact decidable!) set of generative grammars containing a grammar for every decidable stringset and no non-decidable stringset.
- **Theorem 3** There is **no** CE set of grammars containing a grammar for every *infinite* decidable stringset such that every grammar in the set defines an infinite stringset.

The range of our ignorance

In computational complexity we don't know where the proper inclusions are:

$$\underbrace{\text{LogSp} \subseteq \text{P} \subseteq \text{NP} \subseteq \text{Pspace} = \text{NPspace}}_{\text{some proper containments in here}} \subseteq \text{Exp} \subseteq \text{NExp} \subseteq \text{ExpSpace} \cdots$$

In linguistics, we can't decide where English (as a stringset) belongs:

$$\text{SL} \subset \text{LT} \subset \underbrace{\text{SF} \subset \text{FS} \subset \text{LN} \subset \text{DCF} \subset \text{CF} \subset \text{TA} \subset \text{IND}}_{\text{English probably somewhere in here}} \subset \text{CS} \subset \text{PR} \cdots$$

Could English be context-free?

The adverb *respectively*

The actors, admirals, advocates, . . . , and acrobats in Bolton, Birmingham, Bistritz, . . . , and Bilbao are respectively clever, cantankerous, careful, . . . , and curious.

Homomorphic to $\{a^n b^n c^n | n > 0\}$?

No: there is no syntactic constraint here.

??? [NP *Art*], [NP *Bob*], and [NP *Chas*] are married to [NP *Jolene*] and [NP *Karen*] *respectively.*

[NP *The worst recent earthquakes*] *occurred in* [NP *Chile*] and [NP *Japan*] *respectively.*

(Pullum & Gazdar 1982)

Could English be context-free?

Non-identity in comparatives

John was more successful as a biologist$_x$ than he was as a vice chancellor$_y$.

Required non-identity of the nominal strings x and y?

$[_{\text{AdjP}}$ *more* Adjective *as a* $\_\_\_\__x$ *than as a* $\_\_\_\__y$ $]$

No; in the right context, English allows identity:

I'm more successful as a husband than Tiger Woods is as a golfer; in fact right now I'm more successful as a ***golfer*** *than he is as a golfer!*

Moreover, infinitely many stringsets of the form $\{xcy | x, y \in L \wedge x \neq y\}$, where L is CF, are themselves CF (Pullum & Gazdar 1982).

Could English be context-free?

X or no X

We're going ahead, ____ *or no* ____.

Homomorphic to $\{xcx|x \in L\}$, famously non-CF?

No; again the true answer is semantic. And in fact the two strings do not have to be identical:

We're going ahead, stupid management or no stupid bloody management!

The two strings have to be **absolutely identical in sense** (because *X* and *no X* must exhaust all possibilities: Pullum & Rawlins 2007).

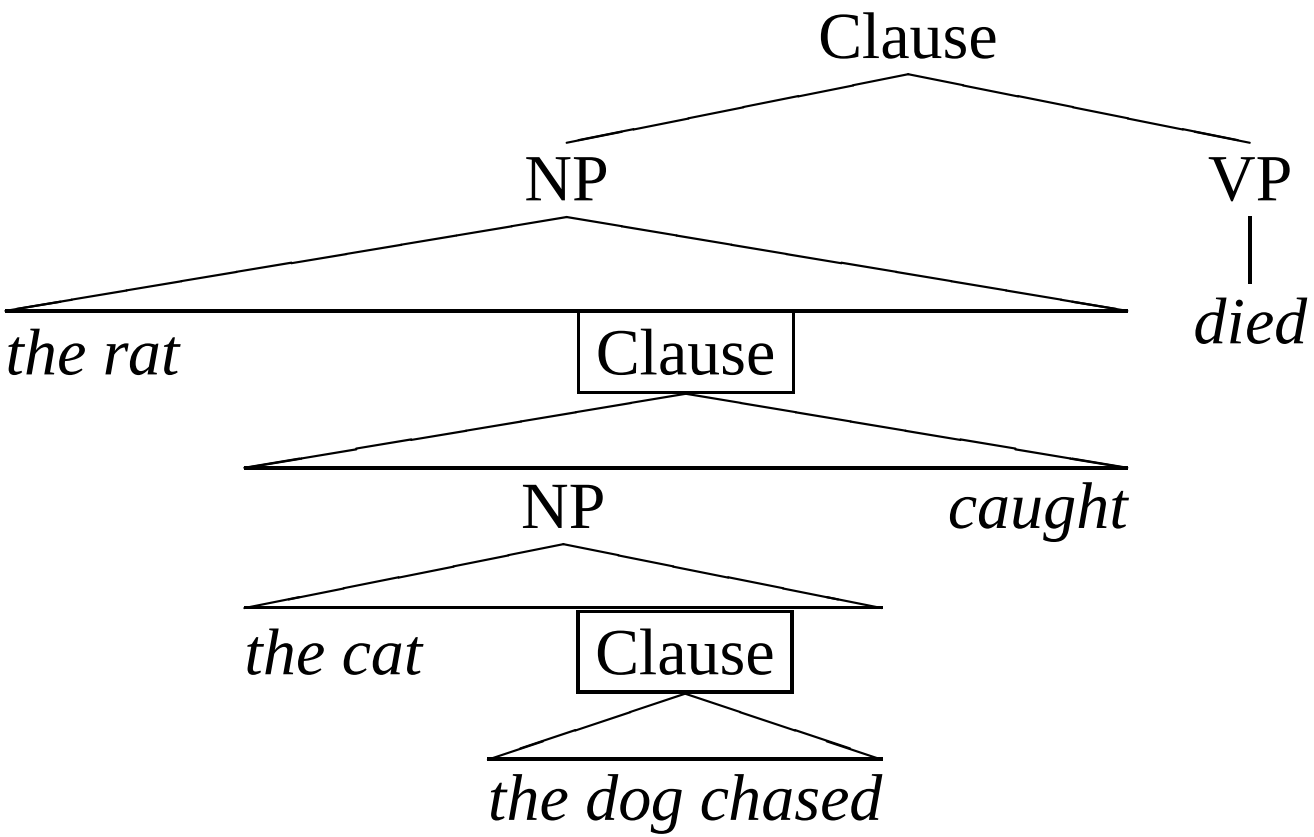
Large number names

How to name a number way bigger than a zillion squared, when *zillion* is the largest number you have a one-word name for:

$$\{\textit{one zillion}^{n_1}\ \textit{one zillion}^{n_2}\ \ldots\ \textit{one zillion}^{n_k} \mid n_i > n_{i+1} \text{ for each } i \text{ such that } 1 \leq i \leq k\}$$

Arnold M. Zwicky (1963) Some languages that are not context-free. *Quarterly Progress Report of the Research Laboratory of Electronics* **70**, 290-293. Cambridge, MA: MIT.

Could English be FS? Center-embedding



Could English be FS?

The rat the cat caught died.	[NP [NP VP]] VP
?*The rat the cat the dog chased caught died.*	[NP [NP^2 VP^2]] VP
??*The rat the cat the dog the bull gored chased caught died.*	[NP [NP^3 VP^3]] VP
???*The rat the cat the dog the bull the vet checked gored chased caught died.*	[NP [NP^4 VP^4]] VP
????*The rat the cat the dog the bull the vet the alligator attacked checked gored chased caught died.*	[NP [NP^5 VP^5]] VP

[. . .]

**The rat squealed died.*	(NP VP^2: too many VPs)
**The rat the cat caught.*	(NP^2 VP: not enough VPs)

Could English be FS?

But all the passives of the rat/cat examples are fully acceptable:

The rat that was caught by the cat died.

The rat that was caught by the cat that was chased by the dog died.

The rat that was caught by the cat that was chased by the dog that was gored by the bull died.

The rat that was caught by the cat that was chased by the dog that was gored by the bull that was checked by the vet died.

The rat that was caught by the cat that was chased by the dog that was gored by the bull that was checked by the vet that was attacked by the alligator died.

[. . .]

Could English be FS?

To argue that English cannot be FS, take English to be a set E containing all of these:

An idiot hired another idiot.

$^{?}$ *An idiot who an idiot had hired hired another idiot.*

$^{???}$ *An idiot who an idiot who an idiot had hired had hired hired another idiot.*

$^{????}$ *An idiot who an idiot who an idiot who an idiot had hired had hired had hired hired another idiot.* [. . . and so on]

Let $R =$ *An idiot* (*who an idiot*)*(*had hired*)* *hired another idiot.*

The intersection of E with R is this set:

$$L = \{\textit{An idiot } (\textit{who an idiot})^n \ (\textit{had hired})^n \textit{ hired another idiot.} \ | n > 0\}$$

But this has the homomorphic image $\{a^n b^n | n > 0\}$, famously not FS.

$E \cap R = L$; R is FS; intersection of FS sets yields FS sets; but L is not FS; therefore (by modus tollens) E is not FS.

Reprise: the range of our ignorance

Again: linguists never arrived at a general agreement concerning where the stringset of English fits:

$$\text{SL} \subset \text{LT} \subset \underbrace{\text{SF} \subset \text{FS} \subset \text{LN} \subset \text{DCF} \subset \text{CF} \subset \text{TA} \subset \text{IND}}_{\text{English probably somewhere in here}} \subset \text{CS} \cdots$$

The question very largely ceased to be under active discussion from the 1990s, despite its importance in principle for computational linguists.

Unnaturalness of human languages as a mathematical class

Is it even sensible to think about the human languages as a stringset class?

It is clear that its properties are mathematically unnatural.

- Closure under homomorphism: the class of human languages cannot possibly be regarded as closed under ‘re-spelling’ of strings.

- Intersection with regular stringsets: the class of human languages cannot possibly be regarded as closed under intersection with regular sets (consider, for example, very small finite ones).

(Observations of Christopher Culy)

Lecture 2: model-theoretic grammar

Plan for a new approach to the syntax of human languages:

- DON'T assume that expressions have to be generated or enumerated, and assigned structures by a rule system
- Assume instead that there already ARE expressions and they HAVE structure
- Take a grammar to be a THEORY in the logician's sense: a set of statements

Model-theoretic grammar

The general idea:

(I) rules are STATEMENTS about expressions;

(II) GRAMMARS are finite sets of such rules;

(III) well-formedness of an expression consists in SATISFACTION of the grammar.

What traditional grammar rules say

This reunites formal linguistics with the policy of traditional grammars in one way. Typical statements:

- 'The subject noun phrase of a tensed clause is in the nominative case'
- 'The main verb of a tensed clause agrees in person and number with the subject of that clause'
- 'Transitive verbs directly precede their direct objects'
- 'Attributive modifiers precede the head words that they modify'

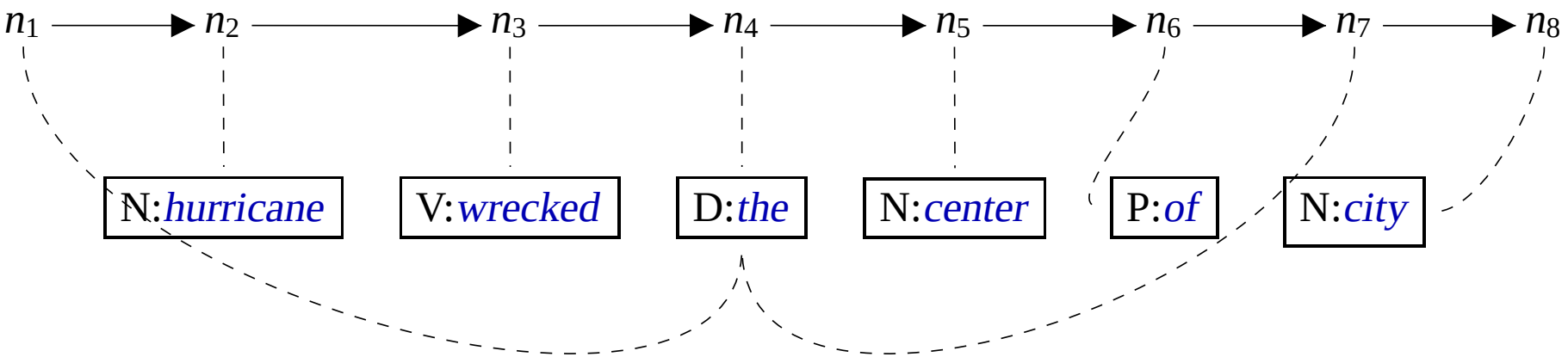
. . .

These are not generative instructions; they are STATEMENTS that are true of properly formed expressions.

A string model

How to model the structure of expressions? A very simple way would be to use strings of categorized words:

The hurricane wrecked the center of the city.



Dashed lines represent predicates of the individuals $n_1, \ldots, n_8$; solid arrows represent a binary strict order on the domain.

Büchi's Theorem

Let $\mathcal{G}$ = regular grammars with some vocabulary V
M = finite string models with points labeled from V
$\mathcal{L}^M$ = a weak monadic second-order (wMSO) language suited to M.

Theorem (Richard Büchi): The following two statements are equivalent:

- $L \approx$ **FinMod**(φ) for some statement φ in a wMSO language suited to description of finite string structures over V.
- L is an FS stringset over V.

Büchi's Theorem

Büchi's Theorem tells us that if (the stringset of) English is FS, then there is a sentence of wMSO that is true of all strings that are grammatical in English and false of all other strings.

Indeed, it can be an existential sentence $(\exists X)[\varphi(X)]$ for FO φ — string models do not suffice to distinguish existential wMSO from all of wMSO.

There is no reference here to generative grammars, or to machines.

The characterization of the FS stringsets is purely in terms of logic.

Model-theoretic characterizations of stringset classes

We can use weaker description languages than wMSO to describe stringsets.

$\mathbf{AP}_k$	=	atomic propositions about k-grams
$\mathbf{PC}_k$	=	propositional calculus on AP_k atoms
$\mathbf{FO}[<]$	=	first-order logic with successor ('immediately precedes')
$\mathbf{FO}[<^*]$	=	first-order logic with less-than ('precedes')
wMSO	=	weak monadic second order logic

Stringset class names:

SL	=	strictly k-local
LT	=	k-locally testable
TT	=	locally threshold-testable
SF	=	star-free (= counter-free)
FS	=	finite-state
CF	=	context-free

Model-theoretic characterizations of stringset classes

A **strictly k-local** description on strings (SL_k^S) over symbol inventory V is a finite set of atomic k-gram propositions consisting of k-length strings over $V \cup \{\blacktriangleright, \blacktriangleleft\}$.

Interpretation:

- atomic proposition **x** means 'the substring x is forbidden';
- atomic proposition **$\blacktriangleright$x** means 'the substring x cannot begin a string'
- atomic proposition **x$\blacktriangleleft$**; means 'the substring x cannot end a string'.

String w is allowed iff $\blacktriangleright w \blacktriangleleft$ has no forbidden k-length substrings.

L is SL_k iff L has an SL_k description, and is SL iff there is any such k.

Model-theoretic characterizations of stringset classes

Example: $\underline{aa^*b^*}$ is described by the $\mathbf{AP}_2$ description $G_1 = \{\blacktriangleright b,\ ba\}$.

G_1 says: (i) a string must never begin with b, and (ii) a string must never have a b followed by an a.

Showing that *aaaabb* is in the set described by G_1:

$\boxed{\blacktriangleright\ a}\ a\ a\ a\ b\ b\ \blacktriangleleft$

$\blacktriangleright\ \boxed{a\ a}\ a\ a\ b\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ \boxed{a\ a}\ a\ b\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ \boxed{a\ a}\ b\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ a\ \boxed{a\ b}\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ a\ a\ \boxed{b\ b}\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ a\ a\ b\ \boxed{b\ \blacktriangleleft}$

Model-theoretic characterizations of stringset classes

Showing that *aaabab* is **not** in the set described by $G_1 = \{\blacktriangleright b,\ ba\}$.

$\boxed{\blacktriangleright\ a}\ a\ a\ b\ a\ b\ \blacktriangleleft$

$\blacktriangleright\ \boxed{a\ a}\ a\ b\ a\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ \boxed{a\ a}\ b\ a\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ \boxed{a\ b}\ a\ b\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ a\ \boxed{b\ a}\ b\ \blacktriangleleft$ ⟵ REJECT

$\blacktriangleright\ a\ a\ a\ b\ \boxed{a\ b}\ \blacktriangleleft$

$\blacktriangleright\ a\ a\ a\ b\ a\ \boxed{b\ \blacktriangleleft}$

Model-theoretic characterizations of stringset classes

Using more and more powerful logics on string models we obtain larger and large classes of stringsets:

$\mathbf{AP}_k^S \quad \rightsquigarrow \quad$ Strictly Local (SL_k) stringsets

$\mathbf{PC}_k^S \quad \rightsquigarrow \quad$ Locally Testable (LT_k) stringsets

$\mathbf{FO}[<]^S \quad \rightsquigarrow \quad$ Locally Threshold Testable (TT) stringsets

$\mathbf{FO}[<^*]^S \quad \rightsquigarrow \quad$ Star-Free (SF) stringsets

$\mathbf{wMSO}^S \quad \rightsquigarrow \quad$ Finite-State (FS) stringsets

And we have an expressive power hierarchy:

$$L(\mathbf{AP}_k^S) \subsetneq L(\mathbf{PC}_k^S) \subsetneq L(\mathbf{FO}[<]^S) \subsetneq L(\mathbf{FO}[<^*]^S) \subsetneq L(\mathbf{wMSO}^S)$$

Or using the abbreviatory names for stringset classes (and adding CF):

$$\mathrm{SL} \subsetneq \mathrm{LT} \subsetneq \mathrm{TT} \subsetneq \mathrm{SF} \subsetneq \mathrm{FS} \subsetneq \mathrm{CF}$$

Moving to tree models

The basic axioms for tree structures can be stated easily in FO:

A1 Connectedness of dominance
$(\exists x)(\forall y)[x \leq y]$
(There is a node that dominates every node, i.e., a root.)

A2 Antisymmetry of dominance
$(\forall x, y)[(x \leq y \wedge y \leq x) \rightarrow (x \approx y)]$
(Two nodes can only dominate each other if they are the same node. This guarantees that the root is unique.)

A3 Transitivity of dominance
$(\forall x, y, z)[(x \leq y \wedge y \leq z) \rightarrow (x \leq z)]$
(Dominating a node entails dominating what it dominates.)

A4 Proper domination (definition)
$(\forall x, y)[(x < y) \Leftrightarrow (x \leq y \wedge \neg(x \approx y))]$
(Defines the '$<$' relation in terms of dominance.)

Moving to tree models

A5 Immediate domination (definition)
$(\forall x, y)[x \triangleleft y \Leftrightarrow (x < y \wedge (\forall z)[(x \leq z \wedge z \leq y) \rightarrow (z \leq x \vee y \leq z)])]$
(Defines '$\triangleleft$' in terms of the dominance relation.)

A6 Discreteness of domination
$(\forall x, z)[x < z \rightarrow ((\exists y)[x \triangleleft y \wedge y \leq z] \wedge (\exists y)[y \triangleleft z])]$
(Every dominance path leading downward from x must include a child of x if it includes any nodes at all other than x.)

A7 Exhaustiveness and Exclusiveness
$(\forall x)(\forall y)[(x \leq y \vee y \leq x) \Leftrightarrow (\neg(x \prec y) \wedge \neg(y \prec x))]$
(Dominance holds, one way or the other, in every pair where precedence doesn't. Thus the union of dominance with precedence and the inverses of both exhausts all the nodes in the tree.)

Moving to tree models

A8 Inheritance of Precedence
$(\forall w)(\forall x)(\forall y)(\forall z)[(x \prec y \wedge x \leq w \wedge y \leq z) \rightarrow \ w \prec z]$
(Preceding a node entails preceding its children.)

A9 Transitivity of Precedence
$(\forall x)(\forall y)(\forall z)[(x \prec y \wedge y \prec z) \rightarrow x \prec z]$
(Preceding a node entails preceding the nodes that it precedes.)

A10 Asymmetry of Precedence
$(\forall x)(\forall y)[x \prec y \rightarrow \neg(x \approx y)]$
(No two nodes precede each other.)

Moving to tree models

A11 Leftmost Child Existence

$(\forall x)[(\exists y)[x \triangleleft y] \rightarrow (\exists y)[x \triangleleft y \wedge (\forall z)[x \triangleleft z \rightarrow \neg(z \prec y)]]]$

(If a node has any children at all, then one of them is the leftmost.)

A12 Discreteness of Precedence

$(\forall x, z)[(x \prec z) \rightarrow (\exists y)[x \prec y \wedge (\forall w)[x \prec w \rightarrow \neg(w \prec y)]] \wedge$
$(\exists y)[y \prec z \wedge (\forall w)[w \prec z \rightarrow \neg(y \prec w)]]]$

(If a node x precedes a node z, then some node is the first one that follows x, and some node is the last one that precedes z. Hence precedence is a discrete ordering like the integers, not a dense one like that of the reals.)

Doner's Theorem

This result emerged partly in consequence of the work of Thatcher and Wright on FS tree automata in the late 1960s:

Theorem (John Doner, 1970): The following two statements are equivalent:

- tree-set $\mathcal{T} \approx$ **Mod**(φ) for some statement φ in a wMSO language suited to description of trees labeled from vocabulary V
- $\mathcal{T}$ is accepted by some finite-state tree automaton using vocabulary V

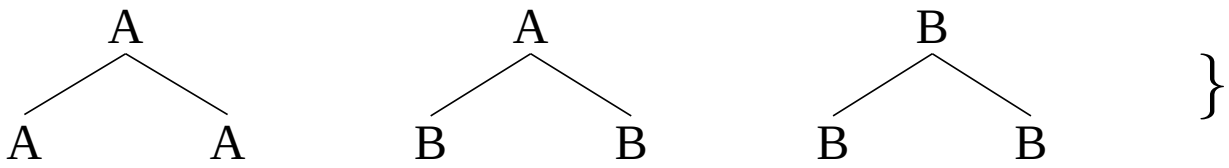
Corollary: If a tree-set $\mathcal{T}$ is **Mod**(φ) for some statement φ in a wMSO language suited to description of trees, then the string yield of $\mathcal{T}$ is CF.

Model-theoretic characterizations of tree-set classes

There are important results about wMSO, but there may be reason to use weaker description languages.

We can use on trees (with the obvious modifications) all the same description languages that we used on strings.

For example, an SL_k description on trees over V is a finite set of atomic V-labeled local trees of depth k. For $k = 2$ and $V = \{A, B\}$, this would be an example:

$G_2 = \{$  $\}$

Interpretation: Each local tree that is one of the atomic propositions is a forbidden subtree.

This description describes the key property of the set of all binary trees in which every node with children has one child (but not both) labeled B.

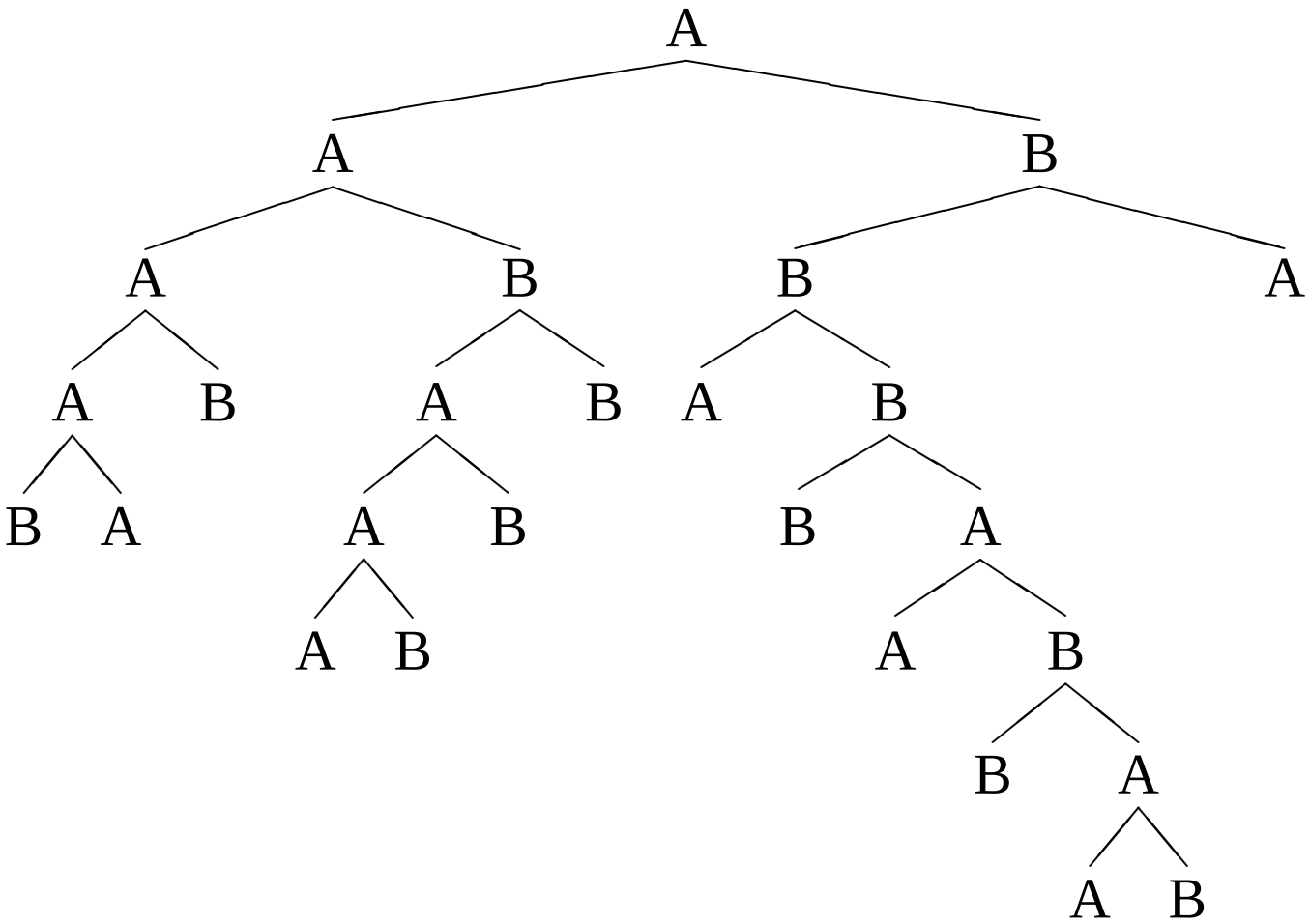
Model-theoretic characterizations of tree-set classes

- A tree is grammatical iff it has no forbidden subtrees.
- A tree-set is SL_k^T iff it has an SL_k^T description.
- A tree-set is SL^T iff there is any such k.

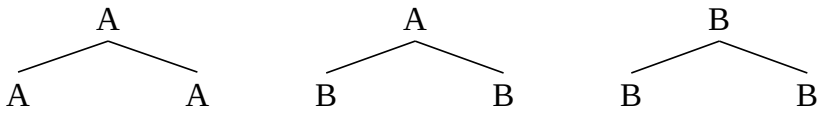
And a tree-set is LT^T iff it can be characterized by some boolean logic expression involving tree-forbidding statements, etc.

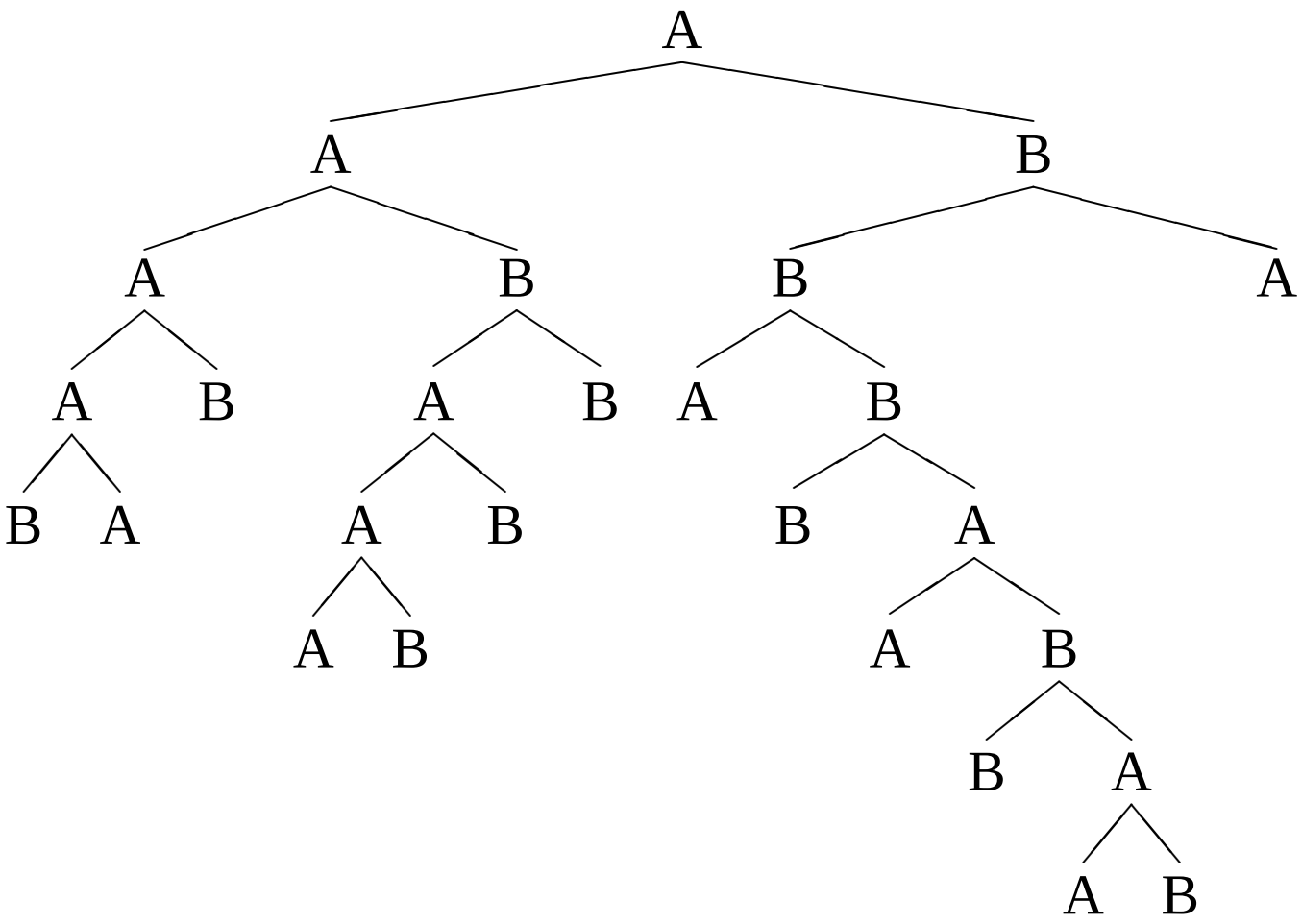
Model-theoretic characterizations of stringset classes

Grammar G_2 (which forbids same-label siblings) allows this tree:



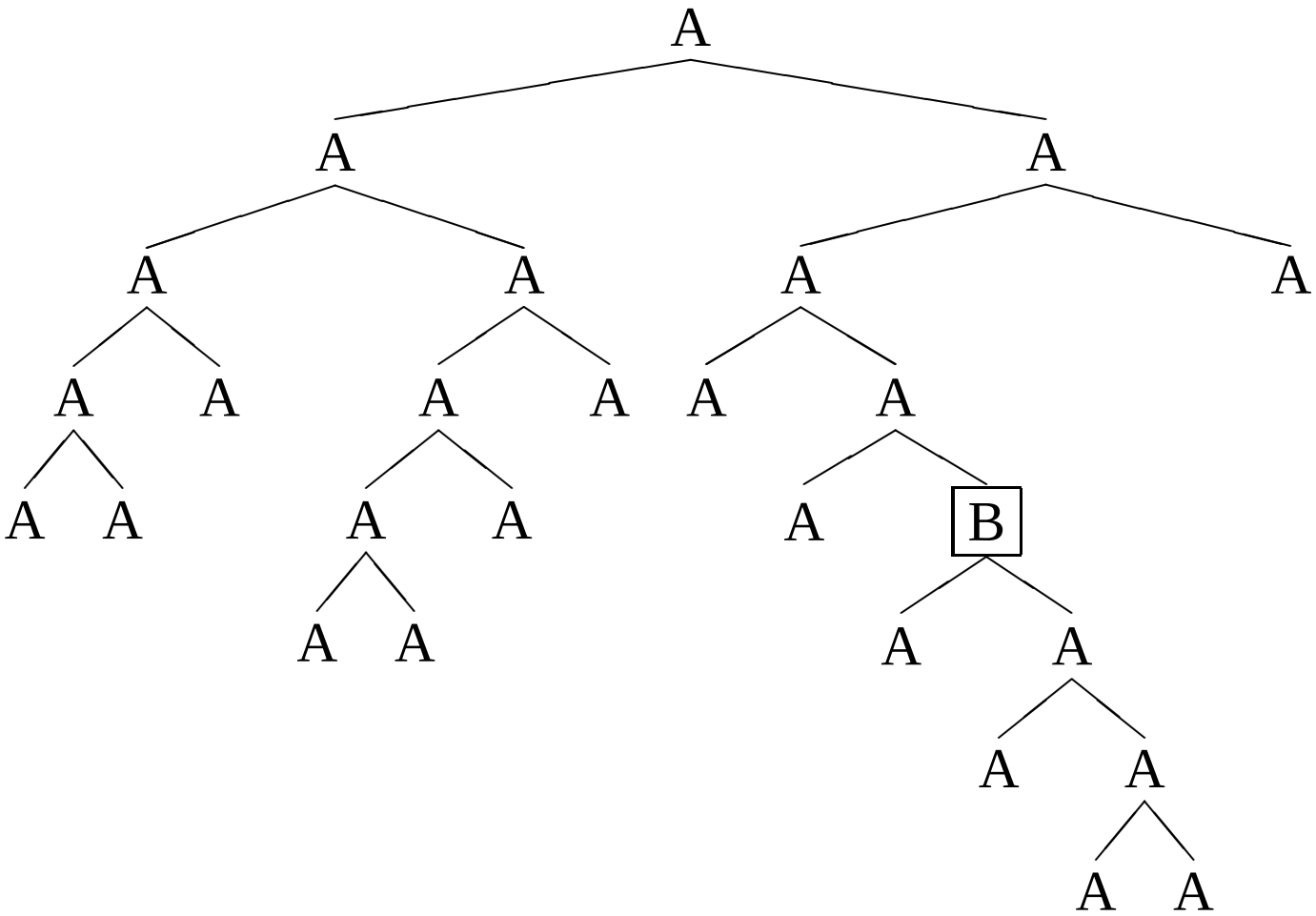
Model-theoretic characterizations of stringset classes

$G_2 = \{$  $\}$ allows this tree:



Model-theoretic characterizations of stringset classes

But now consider the set of all $\{A, B\}$-labeled binary trees containing exactly one B node — trees like this:



Model-theoretic characterizations of tree-set classes

The One-B tree-set cannot be captured by any SL_2^T description.

Indeed, there is no k such that some SL_k^T description can describe it.

The description languages $\mathbf{AP}_n^T$, $\mathbf{PC}_n^T$, $\mathbf{FO}[<]^T$, $\mathbf{FO}[<^*]^T$, and wMSO^T describe progressively larger and larger tree-sets.

For example, a first-order theory including this statement can readily describe the One-B set:

$$(\forall x)[A(x) \underline{\vee} B(x)] \wedge (\exists x)[B(x) \wedge (\forall y)[y \neq x \rightarrow A(x)]]$$

Stringset classes as yields of tree-set classes

We can extract stringsets from tree-sets by taking the string yields of their trees.

Let $\sigma(\mathcal{L}^T)$ denote the string yield obtainable by using the logic $\mathcal{L}$ on trees.

We find a remarkable convergence:

$$\sigma(\mathbf{AP}_n^T) = \sigma(\mathbf{PC}_n^T) = \sigma(\mathbf{FO}[<]^T) = \sigma(\mathbf{FO}[<^*]^T)$$
$$= \sigma(\mathbf{wMSO}^T) = \text{CF}$$

So no matter what logic you use on trees, from $\mathbf{AP}^T$ up to and including wMSO^T, the string yields are the CF stringsets!

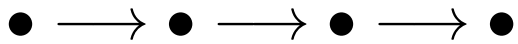
Stringset classes as yields of tree-set classes

DESCRIPTION LANGUAGE	STRING MODELS	TREE MODELS	STRINGSET YIELDS
$\mathbf{AP}_2$	SL_2	2-local tree-sets	CF stringsets
$\mathbf{AP}_3$	SL_3	3-local tree-sets	CF stringsets
$\mathbf{AP}_k$	SL_k	k-local tree-sets	CF stringsets
$\mathbf{PC}_k$	LT_k	LT_k tree-sets	CF stringsets
$\mathbf{FO}(<)$	TT	FO($<$) tree-sets	CF stringsets
$\mathbf{FO}(<^*)$	SF	FO($<^*$) tree-sets	CF stringsets
wMSO	FS	recognizable tree-sets	CF stringsets

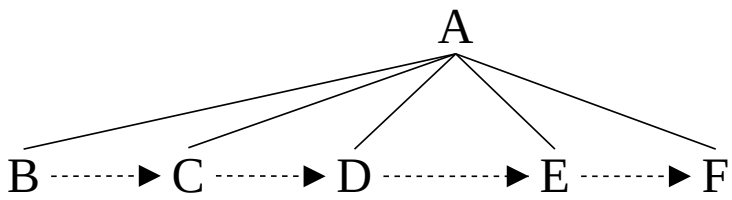
Rogers' Theorem

Dimension 0 (no binary relation): points

Dimension 1 ($\{<_1\}$): strings



Dimension 2 ($\{<_1, <_2\}$): trees



Dimension k ($\{<_1, \ldots, <_k\}$): k-dimensional tree domains

Rogers' Theorem

Theorem (Jim Rogers, 2003): For each $k \geq 0$, wMSO on k-dimensional tree domains defines a distinct class of structures:

$k = 0$	finite stringsets
$k = 1$	FS stringsets
$k = 2$	CF stringsets
$k = 4$	TA stringsets
$\ldots$	

They form an infinite hierarchy, and for each level we have a Myhill–Nerode characterization and a deterministic polynomial recognition result.

Satisfiability of wMSO in the worst case

Satisfiability for wMSO is decidable; but asymptotically it is non-elementary. That is, there is no bound on the repeated exponentiation that might be needed:

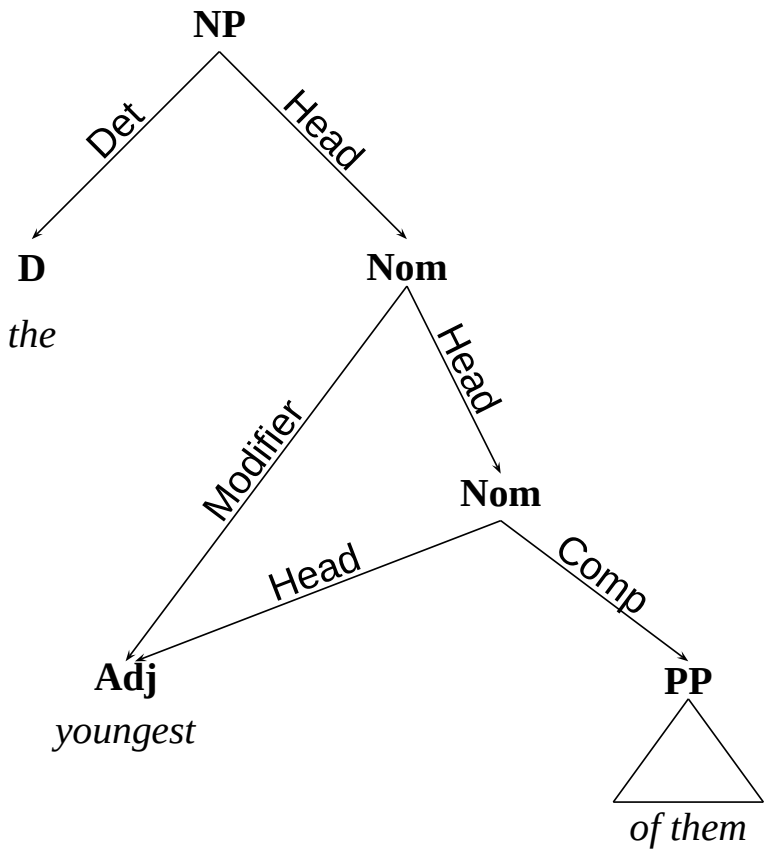
$$2^{2^{\cdot^{\cdot^{\cdot^{2}}}}} \Big\} \text{height of exponent stack} = h$$

The number h depends on the number of quantifier alternations in the formula.

However, this may not matter if quantifiers are only called for in ones and twos.

Non-tree-like structures

Structures do not have to be tree-like structures. *The Cambridge Grammar of the English Language* (Huddleston & Pullum 2002) uses structures such as this:



Non-tree-like structures

Structures with labeled edges, and (especially) with downward convergence of edges (no single-parent condition), are not trees.

But they can be mapped to covering trees by a wMSO-expressible mapping in a way that permits preservation of expressive power results (Pullum & Rogers 2008).

This enables us to say with some confidence (given only the very plausible assumption that nothing said in *CGEL* is inexpressible in wMSO) that the yield of any set of such structures that satisfied the constraints of the grammar will almost certainly be CF.

And we can say this without any formal language-theoretic argument or reference to grammars or automata.

The bottom line

It seems plausible that nearly all of the syntax of a human language might be described by means of logical statements in wMSO or a weaker description language, interpreted on Rogers-style tree-like structures of low dimension:

- perhaps 1 for some languages without any center-embedding;
- probably 2 is reasonable for English;
- conceivably 3 or even 4 for some languages.

Lecture 3: Theoretical implications

Topics to survey:

- the etiology of ill-formedness;
- the status of partially (but not fully) well-formed expressions;
- our robust ability to cope with variation and error;
- expressions containing undefined words;
- the well-formedness of expression fragments;
- the existence of quandary-creating constraint clashes;
- the alleged infinitude of the expressions in a human language.

Etiology of ungrammaticality

A Type 2 rule like this might seem to say that prepositions precede their Noun Phrase complements:

$$\text{PP} \rightarrow \text{P NP}$$

It says no such thing. Suppose either of these rules were also in the grammar:

$$\text{PP} \rightarrow \text{NP P}$$

$$\text{P} \rightarrow e$$

Generative grammars work **holistically** to define a whole set all at once. No part of a generative grammar says anything about any expression.

Gradience of ungrammaticality

A generative grammar for English must generate this:

The growth of spam threatens to make email useless.

And it must not generate this:

**Email growth of make spam the threatens to useless.*

There are no cases other than generating something and not generating it.

An expression is either defined as perfect or not defined at all.

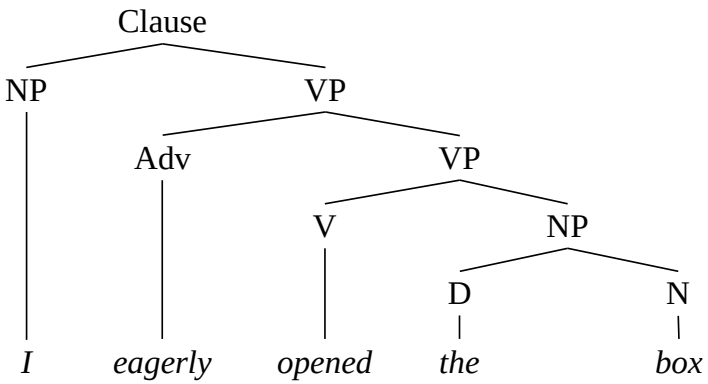
Gradience of ungrammaticality

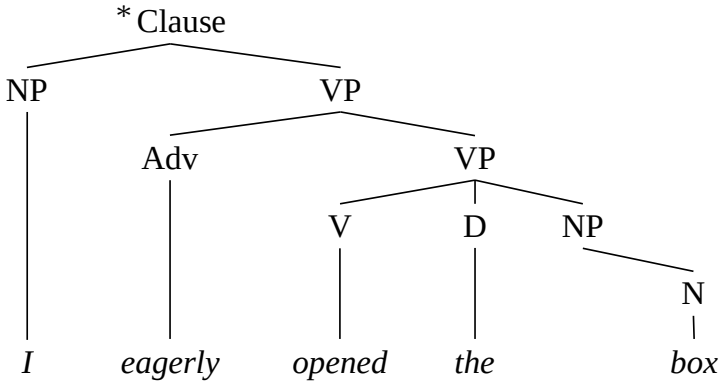
But in fact there are degrees of ill-formedness:

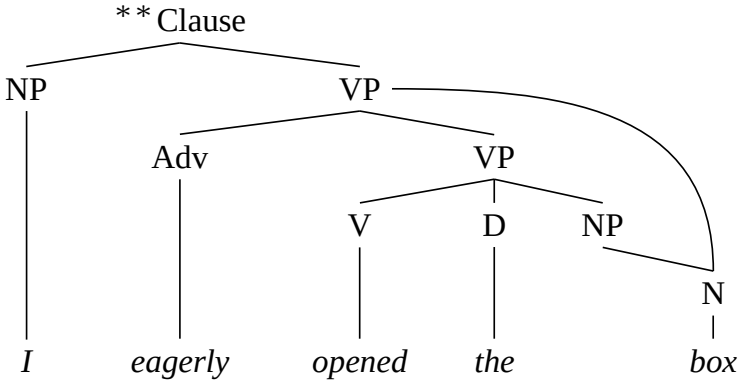
a. *The growth of spam threatens to make email useless.*
b. **The growth of of spam threatens to make email useless.*
c. ***The growth of of the spam threatens to make email useless.*
d. ****The growth of of the spam threatens make email useless.*
e. *****The growth of of the spam threatens make the email useless.*
f. ******The growth of of the spam threatens the email useless make.*
. . .
z. $^{*^{n}}$*The of email growth make threatens spam to useless of.*

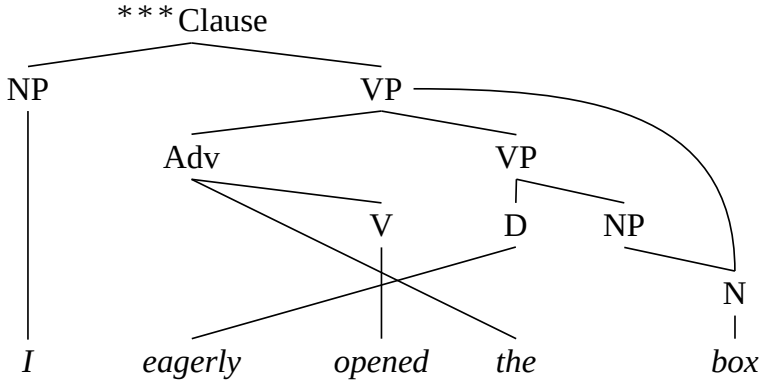
Gradience of ungrammaticality

In principle, it is the same with structures, though the claim that a good string has a bad tree is a rather theoretical one:









Attempting to characterize gradience generatively

a. *John plays golf.*
[animate noun] + [transitive verb needing animate subject and inanimate object] + [inanimate noun] (coarseness level 0)

b. ?*Golf plays John.*
[noun] + [transitive verb] + [noun] (coarseness level 1)

c. **Golf fainted John.*
[noun] + [verb] + [noun] (coarseness level 2)

d. **The of and.*
[word] + [word] + [word] (coarseness level 3)

Attempting to characterize gradience generatively

Input: a string $K_1 \ldots K_n$ of lexical categories corresponding to a string $w_1 \ldots w_n$ of words categorized at some coarseness level $i \leq 3$.

Output: 1 if there is a grammatical sentence that also has lexical category sequence $K_1 \ldots K_n$ at coarseness level i, 0 otherwise.

The main problems:

- nonconstructive definition
- only defines 3 degrees of ungrammaticality
- entirely unrelated to what the grammar does
- and undecidable for transformational grammars!

The model-theoretic representation of ill-formedness

A constraint can be satisfied in a structure at some points (nodes) but not others. For example, the open sentence

$$\mathrm{VP}(x) \rightarrow (\exists y)[x <_2 y \wedge \mathrm{V}(y)]$$

(every VP-labeled node has a child labeled V)

might be true of most VP nodes in a tree but false at one.

And a structure might satisfy nearly all of a set of constraints, but violate just one, perhaps only at one node.

So the notion "almost satisfies $\{\varphi_1, \ldots, \varphi_k\}$ but not quite" is perfectly coherent, provided we take seriously the in-principle separateness of $\varphi_1, \ldots, \varphi_k\}$.

A fine-grained classifications of degrees of ill-formedness is automatically made available by most model-theoretic descriptions.

The model-theoretic representation of ill-formedness

Caution: Such an account will NOT be invariant under reaxiomatization or changes to the vocabulary of category labels.

Restating a grammar in a different form may radically alter its account of degrees of ungrammatically. For example:

- For $\{\varphi_1, \ldots, \varphi_k\}$ (a set of k constraints) there is an upper bound of $(2^k)^n - 1$ ways in which a tree with n nodes could satisfy some of the constraints but not all.
- For $\varphi_1 \wedge \ldots \wedge \varphi_k$ (a single k-conjunct constraint) there are none.

In other words, there may be substantive consequences to the way the linguist decides to formulate a grammar, and in principle facts about degrees of ungrammaticality could be relevant to the choices made.

Robustness

We need to take account of the fact that human languages

(1) nearly always have dialectal variants with slightly different grammar as well as pronunciation, yet people can nearly always understand dialects they do not speak; and

(2) are always used in a way that suffers from occasional errors and slips and idiosyncratic divergences from the norm.

How is it possible that people can understand other dialects and understand people who are making mistakes?

Notice that a generative grammar cannot say anything that helps: it generates just one set of expressions, and says ABSOLUTELY NOTHING about anything outside that set.

Robustness

A model-theoretic description has the potential to make some sense of this robustness in the face of variation and error.

As noted earlier, the notion “You almost respect the constraints on expression structure that I respect, but not quite” is completely coherent.

(Whereas “Your grammar almost generates *x* but not quite” means nothing.)

If 98 percent of my expressions satisfy 98 percent of the constraints you count as defining full membership in your language, that should surely be enough.

Model-theoretic grammars automatically offer at least some hope of accounting for humans’ ability to cope with variation and other people’s errors.

Openness of the lexicon: undefined words

The Gubernator, far ahead in the polls, has several things in his favor.
(From *The Economist*, 4 October 2003, p. 17)

The new Zabundra is even bigger than the Ford Expedition.

"Errrggghh!" went the car as it struggled to get out of the ditch.

All mimsy were the borogoves. (From Lewis Carroll's famous poem 'Jabberwocky'.)

Hand me one of those little cremplefubbers.

In late 3012, the Zorganians attacked the Memphrinons.

My name is Slartybartfast.

Openness of the lexicon: undefined words

But these sentences seem to us not just grammatically well formed, but actually MEANINGFUL.

How could that be explained by a generative grammar that does not generate them — and does not even use a terminal vocabulary to which they belong?

If words are treated as pieces of phonological/orthographic material *constrained by the grammar to have certain syntactic and semantic properties*, then pieces of phonological/orthographic material under no such constraint do not violate any constraint.

Openness of the lexicon: undefined words

Only one thing is wrong when you hear someone tell you to pass him a cremplefubber.

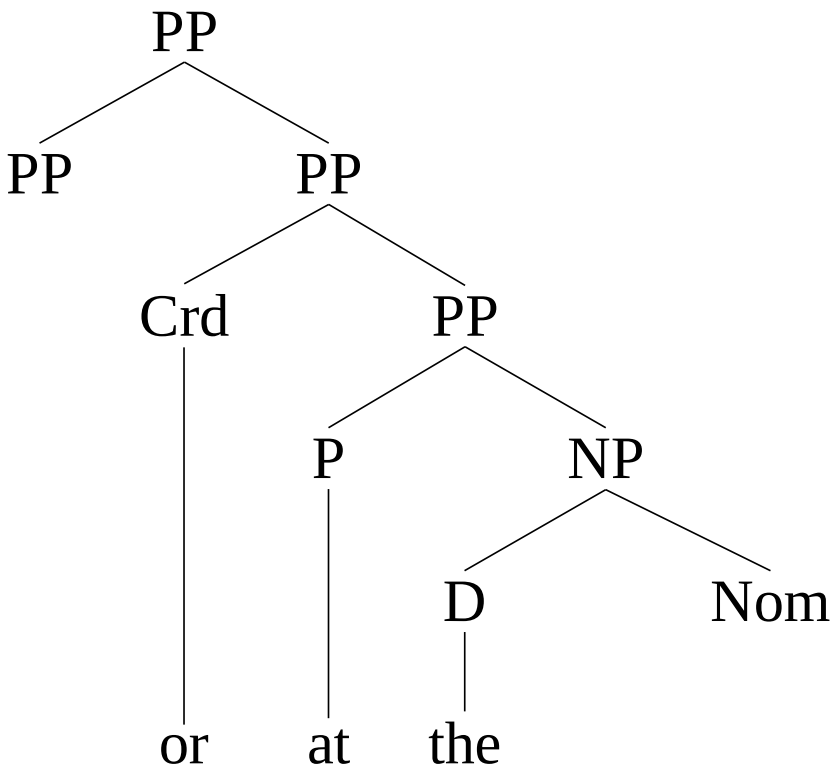
You know what he said, you know what you have to do, so it's not about understanding.

It's just that *you don't know what cremplefubbers are*. That is all.

Nothing about the linguistic structure is amiss, even semantically: *cremplefubber* means roughly what *thing* means, to someone who has never encountered cremplefubbers before.

Fragments

When you hear *or at the*, you hear it as having structure:

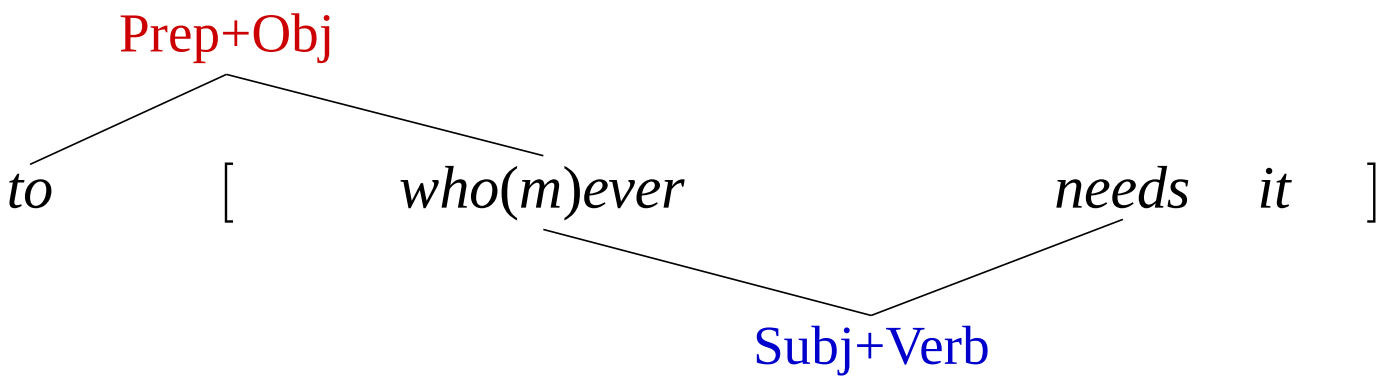


Quandaries

?*I shall give it to whoever needs it.*

?*I shall give it to whomever needs it.*

- A pronoun form must be accusative if it is the head of an NP that is the object of a preposition
- A pronoun form must be nominative if it is the head of an NP that is the subject of a finite verb.



Infinitude

Strange recent remarks by linguists (1)

> "Infinity is one of the most fundamental properties of human languages, maybe the most fundamental one. People debate what the true universals of language are, but indisputably, infinity is central."

(Howard Lasnik, 2000)

Infinitude

Strange recent remarks by linguists (3)

> "This property of discrete infinity characterizes EVERY human language; none consists of a finite set of sentences. The unchanged central goal of linguistic theory over the last fifty years has been and remains to give a precise, formal characterization of this property and then to explain how humans develop (or grow) and use discretely infinite linguistic systems."

(Sam Epstein and Norbert Hornstein, 2005)

Infinitude

Strange recent remarks by linguists (3)

> "[M]any have argued that the property of recursive infinity is perhaps *the* defining feature of our gift for language."
>
> (Charles Yang, 2006)

Infinitude

The supposed inductive argument for the claim that English has infinitely many grammatical expressions:

— *very nice* is grammatical

— adding one *very* makes *very very nice*, which is grammatical

— adding another *very* makes *very very very nice*, which is grammatical.

$\cdots$ (and so on) $\cdots$

— So by induction, for every natural number n, adding one extra *very* to *very*^{n} *nice* makes an expression *very*$^{n+1}$ *nice* which is also grammatical.

Infinitude

But “for every natural number *n*” gives the game away: the question has been begged.

The decision that induction on the natural numbers can be used in this doman has ALREADY PRESUPPOSED that the domain is infinite.

On domains where we know the infinitude conclusion cannot be correct, we simply reject the appropriateness of the reasoning.

For example . . .

Infinitude

A stupid argument in human biology:

- 1 year is a biologically possible age for humans.
- Adding one year of life to a human of age 1 gives an age of 2, which is also biologically possible.
- Adding one further year of life gives an age of 3, which is also biologically possible.

  $\cdots$ (and so on) $\cdots$

- So by induction, for every natural number n, adding one extra year of life to a human of age n gives an age of $n + 1$, which is also biologically possible.

(Conclusion false because of the Hayflick limit.)

Infinitude

A stupid argument in evolutionary biology:

- This organism is of the species Canus lupus familiaris.
- Its female ancestor one generation back was a female organism also of the species Canis lupus familiaris.
- Its female ancestor one generation before that was a female organism also of the species Canis lupus familiaris.

  $\cdots$ (and so on) $\cdots$

- So by induction, for every natural number n, at n generations back its ancestor $n + 1$ generations back was a female organism also of the species Canis lupus familiaris.

(Conclusion false because dogs were only domesticated from the gray wolf about 15,000 years ago.)

Infinitude

We have to ask how we know that the argument used for the claim that English has infinitely many sentences is a sensible one, not one of the many stupid ones.

We need grounds for claiming (a):

(a) extension in sentence length and complexity goes on forever without altering grammaticality

rather than claiming (b):

(b) extension in sentence length tapers off gradually and ceases to preserve grammaticality after some (rather vaguely defined) point is reached.

We don't have any such non-question-begging grounds.

Infinitude

Even in pure mathematics we know of cases where a long succession of cases where some claim is true can be followed by infinitely many more where it is false.

Take the prime-counting function $\pi(x)$ and the logarithmic integral function $\mathrm{li}(x)$.

It has been shown computationally by Kotnik (2008) that there are no values of x below 10^{14} for which $\pi(x) > \mathrm{li}(x)$.

Yet Stanley Skewes proved long ago that eventually there are values of x where $\pi(x) > \mathrm{li}(x)$ (in fact there are infinitely many crossing points).

So where are the calculations by linguists on the matter of maximum expression complexity? There aren't any.

Infinitude

Or rather, when evidence is gathered or calculations are done, linguists tend to ignore both.

- Fred Karlsson searched carefully for sentences with significant depths of initial or center-embedding, and found hardly anything.

  But linguists continue to believe what they believed before: that initial embedding and center-embedding to any degree are grammatical and the set of sentences exhibiting them is **infinite**.

- András Kornai did some statistical analysis on the frequencies of attested words and showed that the data clearly have the profile you would expect from an **infinite** population of words.

  But linguists continue to believe what they believed before: that the set of words is **finite**.

Infinitude

What consequences flow from the supposed infinite number of sentences in human languages?

None.

- Nothing follows about use of the language
- No theoretical claims build interestingly upon it
- No evidence directly confirms it.
- No evidence refutes it, or ever could.

Infinitude

Only one suggestion has much plausibility.

A generative grammar for a large finite set of expressions is very tedious to construct. Walter Savitch has shown that infinitely many finite stringsets have infinite extensions with exponentially shorter grammars.

Recursive rule application is the obvious solution to many descriptive problems. And where there is non-trivial recursive rule application, a generative grammar will generate infinitely many strings (the cases where this does not happen can be regarded as somewhat pathological).

If we assume linguists have mistaken the effects of their descriptive technology for a property of their subject matter, we have an explanation for their otherwise strange infatuation with infinitude.

Infinitude

If they are not simply being misled by generative grammars, we need to ask why linguists cling to the belief that human languages have infinitely many expressions when

(i) it may well be false of some languages (e.g., Pirahã), and

(ii) it is empirically unsupported and unsupportable even for English, and

(iii) if true it would make no difference.

They may feel infinitude is closely tied to the **creativeness** of language use: People make up, utter, and understand sentences that have never been encountered before.

But connecting creativity to infinity is a mistake. Think of (i) chess, (ii) bridge, or (iii) composing sonnets or *haiku.*

Infinitude

The connection to an implication of model-theoretic syntax is very straightforward.

How many graphs are there that satisfy the transitivity condition $(\forall x, y, z)[E(x, y) \wedge E(y, z) \rightarrow E(x, z)]$?

As many as you want to say there are. Given a finite class of finite candidate models (say, the set of graphs representing sets of human beings who know each other), it is some finite number.

Given the class of all finite graphs as candidates, is countably infinite (though vanishingly small asymptotically as a proportion: as larger and larger randomly constructed graphs are considered, the probability of a graph satisfying transitivity falls away to become zero in the limit).

Infinitude

Just so with linguistic expressions. If English has just finitely many expressions and they are of finite size, then only finitely many structures will satisfy the grammar.

If there is no limitation to a finite number, then perhaps infinitely many satisfy the grammar.

The rules of the grammar, the syntactic constraints, will be the same in either case.

We do not need to stipulate an answer: *we can describe syntactic structure in a way that does not entail any commitment regarding how many expressions exibit that structure*.

THANK YOU!

I have enjoyed being here in Sofia
and giving these lectures.

Geoffrey K. Pullum • **gpullum@ling.ed.ac.uk**
`http://ling.ed.ac.uk/~gpullum`

School of Philosophy, Psychology, and Language Sciences
University of Edinburgh
Dugald Stewart Building
3 Charles Street
Edinburgh EH8 9AD, UK