Applying Logic to the Study of Human Language Syntax

Geoffrey K. Pullum

Brown University and the University of Edinburgh

August 2012
East Asian School in Logic, Language and Computation

Lecture 2: Generative Grammars

Recapitulation

What we have done so far

- Human languages have indefinitely many expressions, and differ radically from each other in syntax.
- Emil Leon Post's work laid the foundations for GES grammars, which describe languages via rules ('productions') which are very much like rules of inference in logic.
- GES grammars can generate any set of strings that has any finite specification of membership at all (= any CE set).
- This remains true under radical simplifications of rule form.
- Generative power can be cut by tighter restrictions on rule form, guaranteeing decidability of the generated **stringsets** (= sets of strings).
- Four proper subsets of CE are particularly well known:
  CE $\supset$ CS $\supset$ CF $\supset$ FS $\supset$ Finite

Transformations

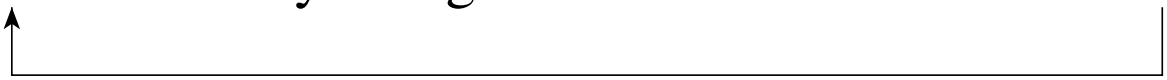
Noam Chomsky suggest using **transformational grammars** for human languages.

A transformational grammar consists of a CF grammar plus a set of **transformational rules** that operate on the objects it produces.

One example will serve to illustrate. . .

Transformations

I wonder which room they thought we had moved to ___?



Chomsky's notation:

T_{w_1}: Structural analysis: $X - NP - Y$ (X or Y may be null)

Structural change: $X_1 - X_2 - X_3 \rightarrow X_2 - X_1 - X_3$

Post's notation:

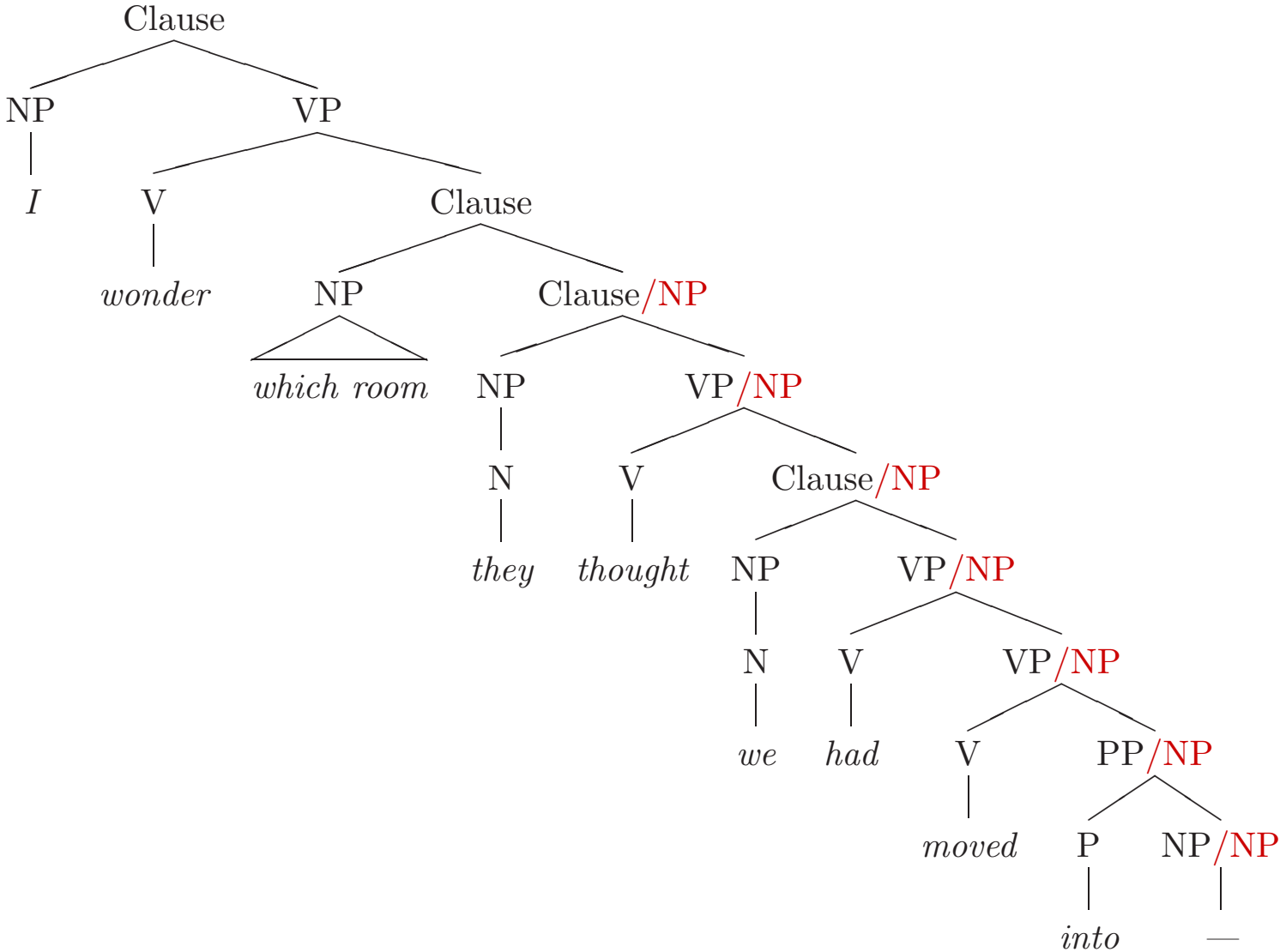
$P_1\ NP\ P_2$ produces $NP\ P_1\ P_2$

A remark by Hilary Putnam (1961):

> 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

Transformations are not needed if V_N is expanded (Gazdar 1981):



Proper subsets of the FS class

What classes of stringsets are properly included within FS (Finite State = Regular) ?

Type 0 (CE) $\supseteq$

Type 1 (CS) $\supseteq$

Type 2 (CF) $\supseteq$

Type 3 (FS) $\supseteq$

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

? ? ? $\longleftarrow$ WHAT CLASSES LIE IN HERE?

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

Finite stringsets

Proper subsets of the FS class

Finite State (FS)

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

Star Free (SF)

Locally Testable (LT)

Strictly Local (SL)

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

Finite stringsets

$FS \supset SF \supset LT \supset SL \supset FINITE$

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.

An LT description can 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.

An SF description can 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.

Reminder: regular expressions

a	a
ab	ab
a+b	$\{a\} \cup \{b\}$
a*	$\{a, aa, aaa, aaaa, \ldots\}$
a*+b*	$\{a, aa, aaa, aaaa, \ldots\} \cup \{b, bb, bbb, bbbb, \ldots\}$
$\overline{\mathbf{a}}$	all strings other than a
$\overline{\mathbf{a}*}$	all strings not in $\{a, aa, aaa, aaaa, \ldots\}$

And where ϕ and ψ are regular expressions denoting $d(\phi)$ and $d(\psi)$, and we write aa as a^2, aaa as a^3, etc.:

ϕ**+**ψ	$d(\phi) \cup d(\psi)$
$\phi\psi$	$\{xy \mid x = d(\phi) \text{ and } y = d(\psi)\}$
ϕ*	$\{x \mid x = x_1 \ldots x_n \text{ and } \forall i_{(0 \leq i \leq n)} [x_i \in d(\phi)]$

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: $\overline{L} =_{\text{def}} V^* - L$.

Example: $a^*ba^*ca^*$ 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 containing neither b nor c, followed by c, followed by anything containing neither b nor c'

Proper supersets of CF

Type 0 (CE) $\supseteq$

Type 1 (CS) $\supseteq$

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

? ? ? ⟵ WHAT CLASSES LIE IN HERE?

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

Type 2 (CF) $\supseteq$

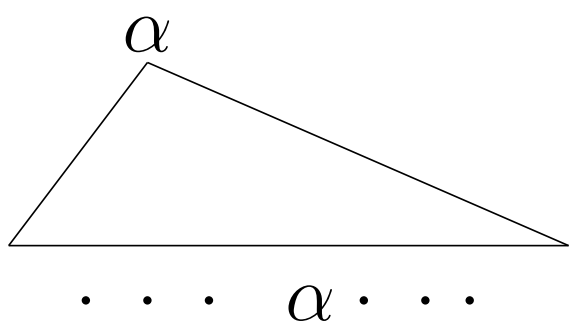
Type 3 (FS) $\supseteq$

Subregular classes

Finite stringsets

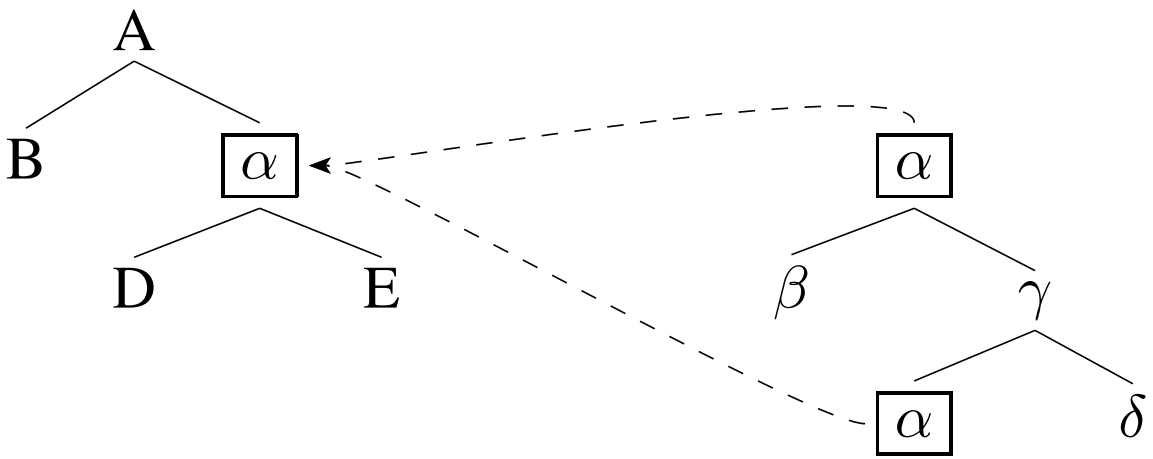
Tree Adjoining (TA) stringsets

An auxiliary tree:



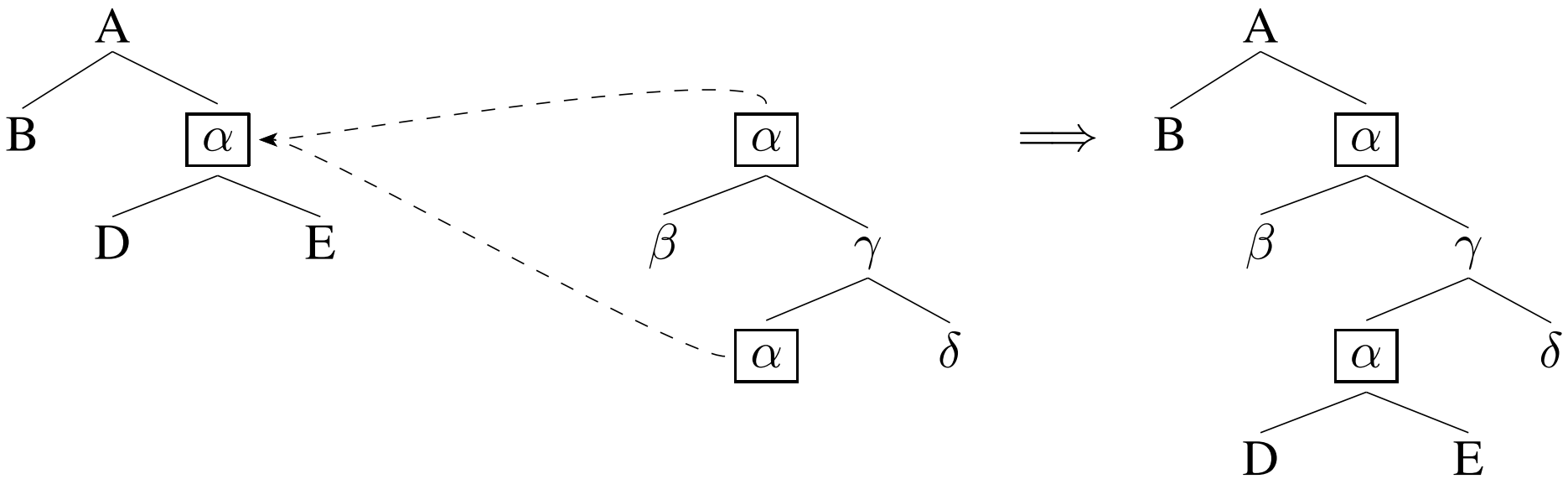
Tree Adjoining (TA) stringsets

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 (Pollard, 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?)

The range of our ignorance

Just as in computational complexity we don't know where the proper inclusions are:

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

. . . similarly in linguistics, we can't decide where human languages (considered as stringsets) might belong:

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

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 $\{xcy | x, y \in L \wedge x \neq y\}$, 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).

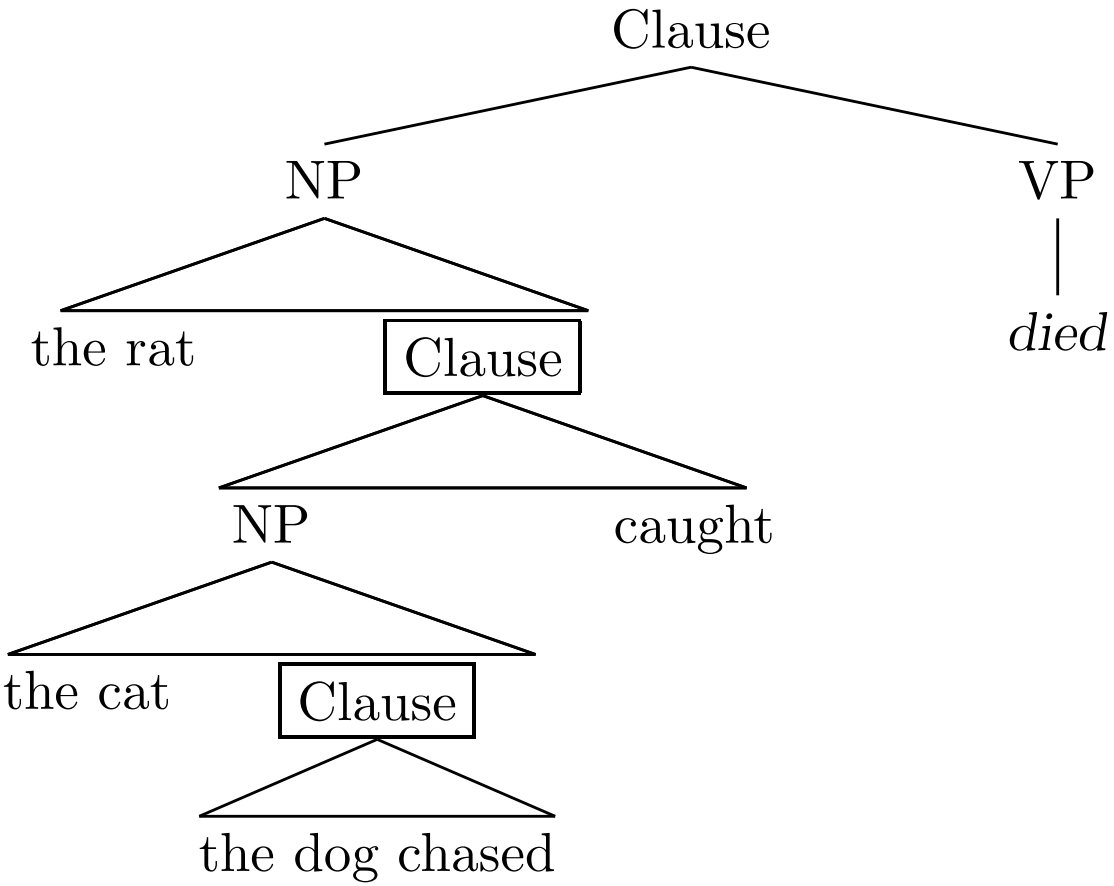
Large number names

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:

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

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)

Alternatives to generative grammars

There are ways in which explicit grammars could be stated without being production systems.

Grammars could be transducers (mappings from expressions to expressions). [Manaster-Ramer; Shieber; . . .]

Grammars could be collections of operations on collections as in category theory.

Grammars could be sets of statements about structure given in an interpreted logical metalanguage.

Our next step is to explore the third of these alternatives.