What makes a language a human language? Mathematical perspectives on Universal Grammar

Meaghan Fowlie

Utrecht Institute of Linguistics OTS

Connecting Mathematical Methods across Utrecht University
Centre for Complex Systems Studies
May 3, 2019

What is linguistics?

The scientific study of human language as a natural phenomenon

Mysteries

Your language lives in your brain. How did it get there? What does it mean to know a language?

Mystery: language acquisition

- Learns her native language in ca. 5 years
- Can't hear all ∞ sentences
- Can't reason

Innateness hypothesis:
She is born ready for (or knowing?) language

Universal Grammar Hypothesis

All human languages share a higher level grammar, with linguistic variation fairly minor variants within the universal grammar

- What's in UG?
- How does she use it to learn her particular language?
  Attempts to characterise UG usually fail when they are specific:
- All languages have nouns, verbs, adjectives, adverbs, and prepositions? Nope
- All languages have reflexive pronouns like *herself* that can only occur in restrictive contexts (eg not at the start of a sentence)? Nope.

  (1) Herself is attending the gala (Scottish English)

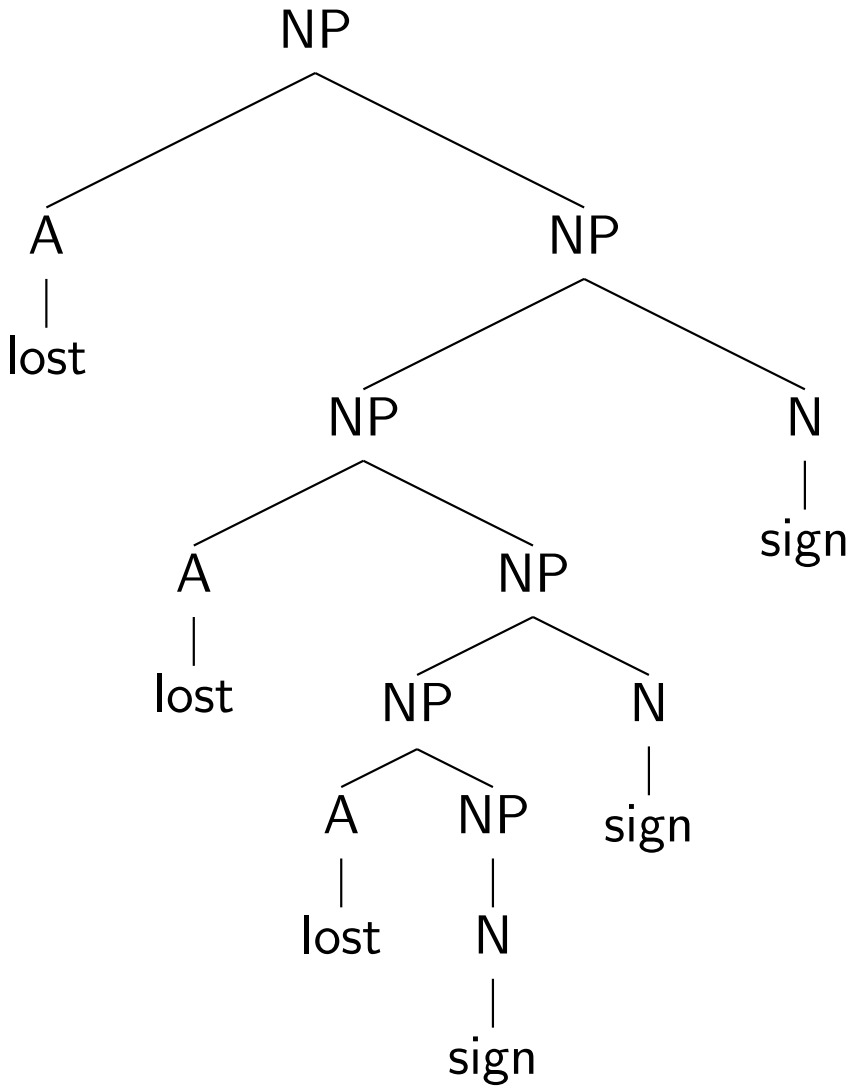
- All languages have recursion? Maybe (Hauser et al., 2002)

Recursion: infinite use of finite means

Phrases can appear inside the same kinds of phrases

Recursion: infinite use of finite means

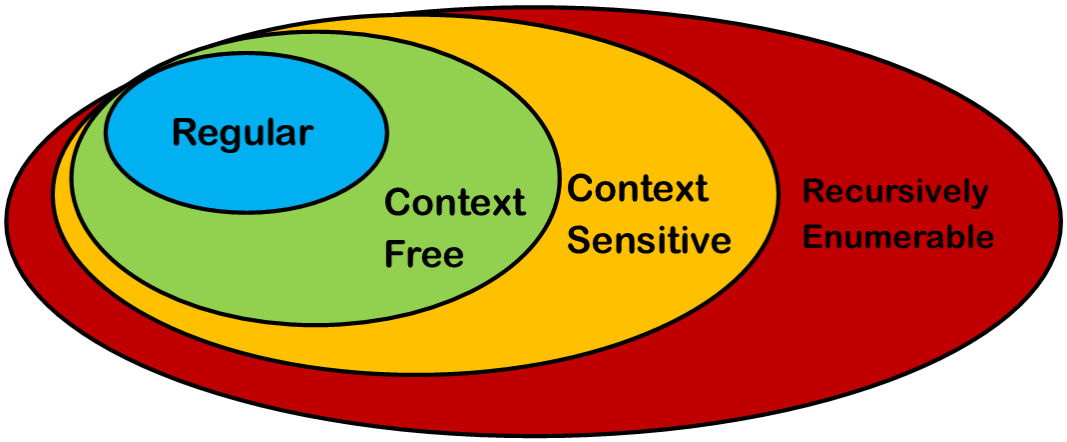
Phrases can appear inside the same kinds of phrases



Recursion

- Pirahã (Brazil) might not have recursion! Everett (2009)
- BUT this is a more promising avenue: it is more mathematical.
- Maybe we should say rather

  all human languages have the kind of grammar that can have recursion



Phonology

- The sound system of a language
- There are rules!

ngayón – pronounced /ŋajon/

- Tagalog, meaning ‘now’
- **impossible English or Dutch word** because it violates the rules of English and Dutch phonotactics.

English and Dutch can’t have ŋ at the start of a word

How to learn the phonotactic rules of English (Try 1)

1. Take all the existing words of English,
2. Put them in IPA, and add symbols for the start and end of the word and between syllables
3. Make a list of every pair of symbols that occur next to each other
4. That's it! You have a *Strictly Local grammar* for English phonotactics.

E.g.:
phonology $\rightarrow$ /⋊fə.nɑ.lə.dʒi⋉/ $\rightarrow$ ⋊f fə ə. .n nɑ ɑ. .l lə ə. .dʒ dʒi i⋉

/fə.dʒi/, /fə.lə.lə.nɑ.dʒi/, /fə.nɑ.nɑ.lə.nɑ.dʒi/

Strictly Local Languages

Strictly Local language: Set **B** of all **legal bigrams** in the language. All finite words with only legal bigrams are in the language

- Σ: finite set of symbols (the *alphabet*)
- Σ^*: all finite sequences of symbols from Σ.
- The *language of B* (L(B)):

$$\{w \in \rtimes \Sigma^* \ltimes \mid \forall i < |w| \quad w_i w_{i+1} \in B\}$$

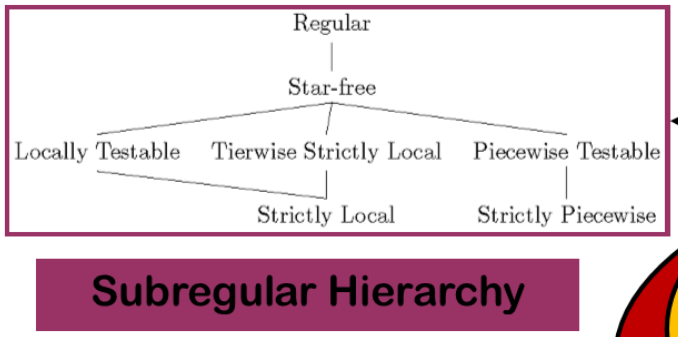
- (Not crazy but also not true) hypothesis: the possible words of a human language are always a Strictly Local language

Learning phonotactics

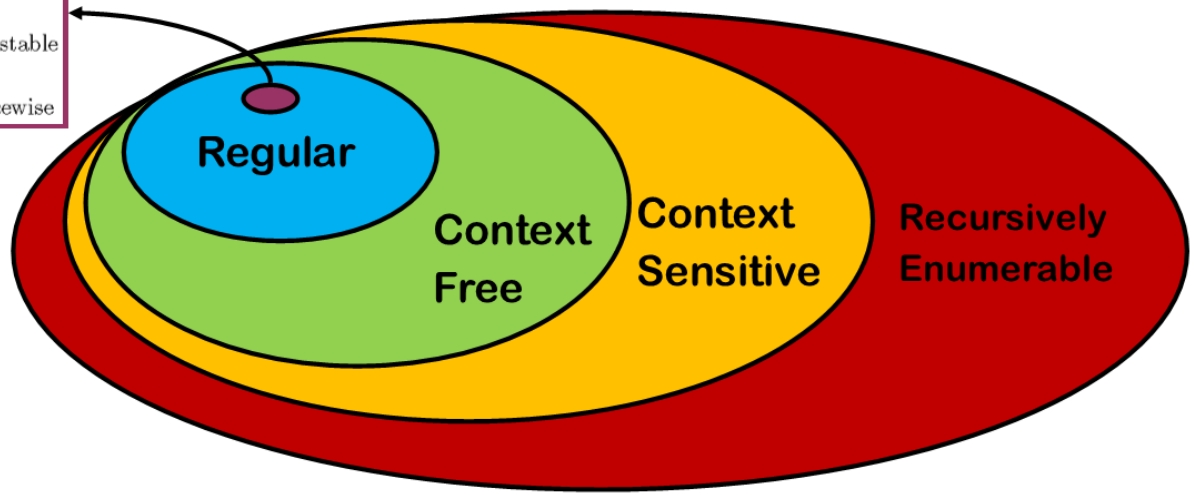
- Learnable: **if the baby knows she should**, she can listen for and memorise all bigrams she hears (frequently enough)
- Then she can learn the Strictly Local phonotactics of her langauge
- "Knowing you should" is part of UG

Infinite use of finite means

$B = \{\rtimes a, aa, a \ltimes\}$ $L(B) = a^*$ Infinite language!



Subregular Hierarchy



Tier-based Strictly Local

"Yurok has a rhotic vowel harmony process by which underlying non-high vowels /a/, /e/, and /o/ may become /ə˞/ in a word that has /ə˞/; for example, the root /nahks-/ 'three' becomes [nə˞hks-] in the word [nə˞hksə˞ʔə˞jɬ] 'three (animals or birds)'."

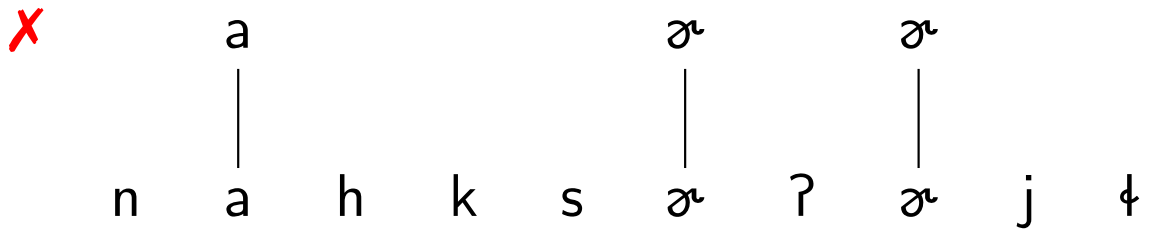
-Survey of California and Other Indian Languages at UC Berkeley

Tier-based Strictly Local

(2) nahks + -ɚʔɚjɬ = nɚhksɚʔɚjɬ
three + ANIMALS = three.ANIMALS
'three (for animals)'

Vowel tier: Look just at the vowels, and ignore the consonants
Illegal bigrams:

✗ aɚ, ɚa, eɚ, ɚe, oɚ, ɚo

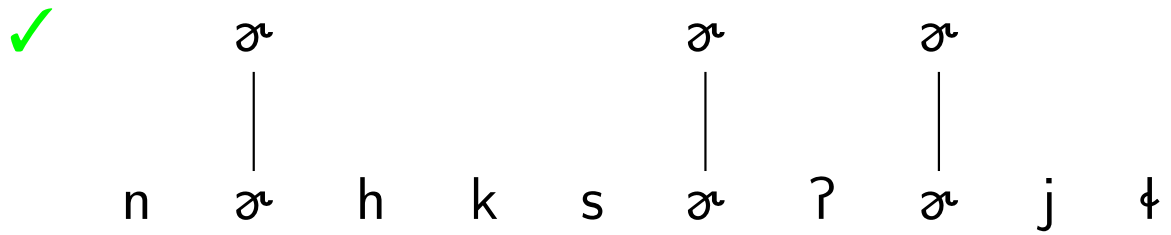


Tier-based Strictly Local

(3) nahks + -ɚʔɚjɬ = nɚhksɚʔɚjɬ
three + ANIMALS = three.ANIMALS
'three (for animals)'

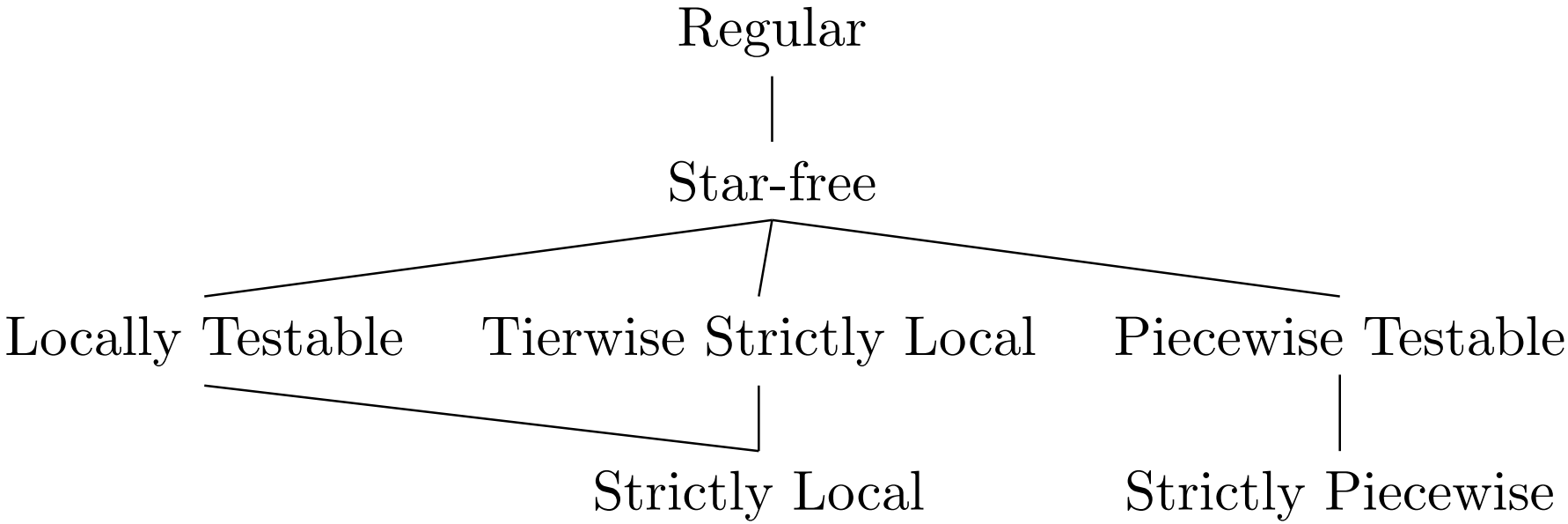
Vowel tier: Look just at the vowels, and ignore the consonants
Illegal bigrams:

✗ aɚ, ɚa, eɚ, ɚe, oɚ, ɚo



Subregular hierarchy

Hypothesis (Heinz et al., 2011) : all human language *phonotactics* (rules about the possible words of a language) can be described by a Tier-based Strictly Local grammar



TSL learnable too (Jardine and Heinz, 2016)

Syntax – the sentences of a language

Can we use subregular grammars to describe human language syntax? Is human syntax subregular?

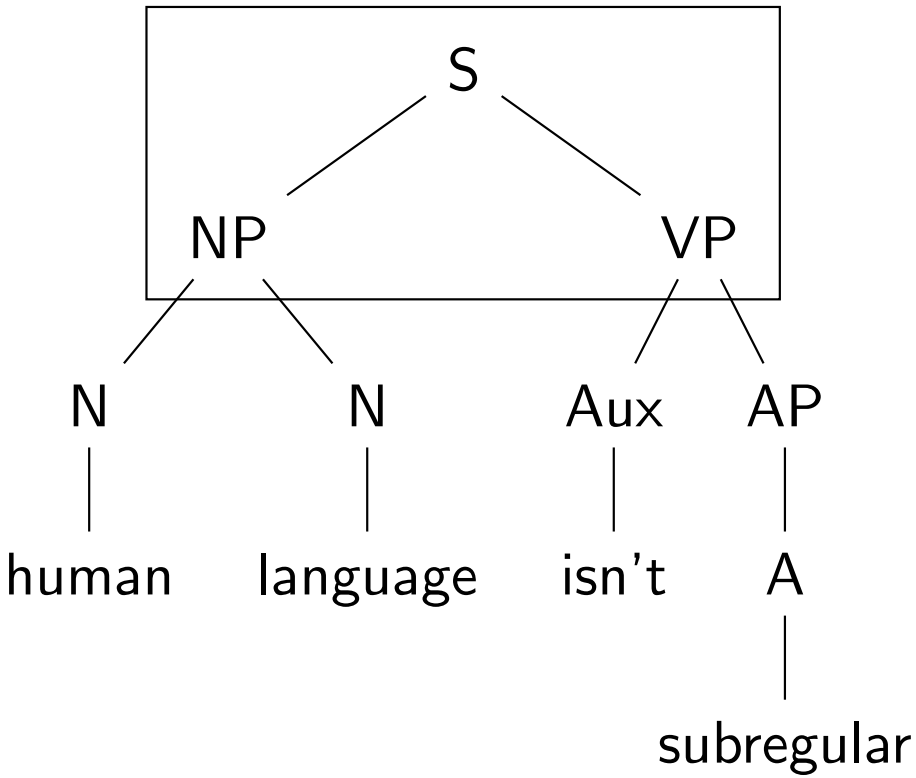
$\{\rtimes$ Can, Can we, we use, use subregular, subregular grammars, grammars to, to describe, describe human, human language, language syntax, syntax $\ltimes$, $\rtimes$ is, is human, human syntax, syntax subregular, subregular $\ltimes\}$

$\boxed{\rtimes \text{ is}}$ human language syntax subregular grammars to describe human syntax subregular $\ltimes$

$\rightarrow$ definitely not Strictly Local (bigrams)!

(Not TSL either. nor Regular)

Syntax



$$
\begin{aligned}
S &\rightarrow \text{NP VP} \\
NP &\rightarrow \text{N N} \\
VP &\rightarrow \text{Aux AP} \\
AP &\rightarrow \text{A} \\
N &\rightarrow \text{human} \\
N &\rightarrow \text{language} \\
Aux &\rightarrow \text{isn't} \\
A &\rightarrow \text{regular}
\end{aligned}
$$

Strictly Local **tree** languages

let $G = \langle B, S, \Sigma \rangle$
B: "tree bigrams",
S: set of Start symbols,
Σ: set of Terminal symbols

eg:
$S = \{S\}$
$\Sigma = \{$human, language, isn't, regular$\}$

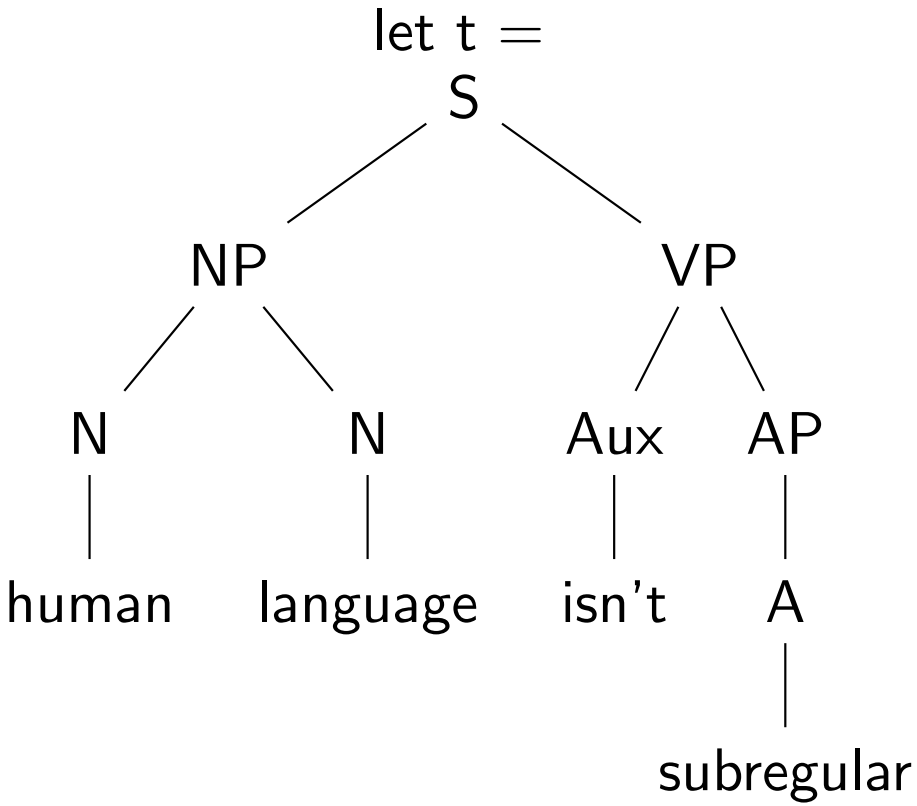
eg B=

S	$\rightarrow$	NP VP
NP	$\rightarrow$	N N
VP	$\rightarrow$	Aux AP
AP	$\rightarrow$	A
N	$\rightarrow$	human
N	$\rightarrow$	language
Aux	$\rightarrow$	isn't
A	$\rightarrow$	regular

tree language of G $L_T(G)$ = set of all trees all of whose "bigrams" are in B, and whose root is in S and whose leaves are in Σ.

Compare $\quad \{w \in \rtimes \Sigma^* \ltimes \mid \forall i < |w| \quad w_1 w_{i+1} \in B\}$

String yield



The *string yield* of a tree is all of its leaves, in order

yield(t) = *human language isn't subregular*

Context Free Grammars

String language of G

$$L_s(G) = \{yield(t) \mid t \in L_T(G)\}$$

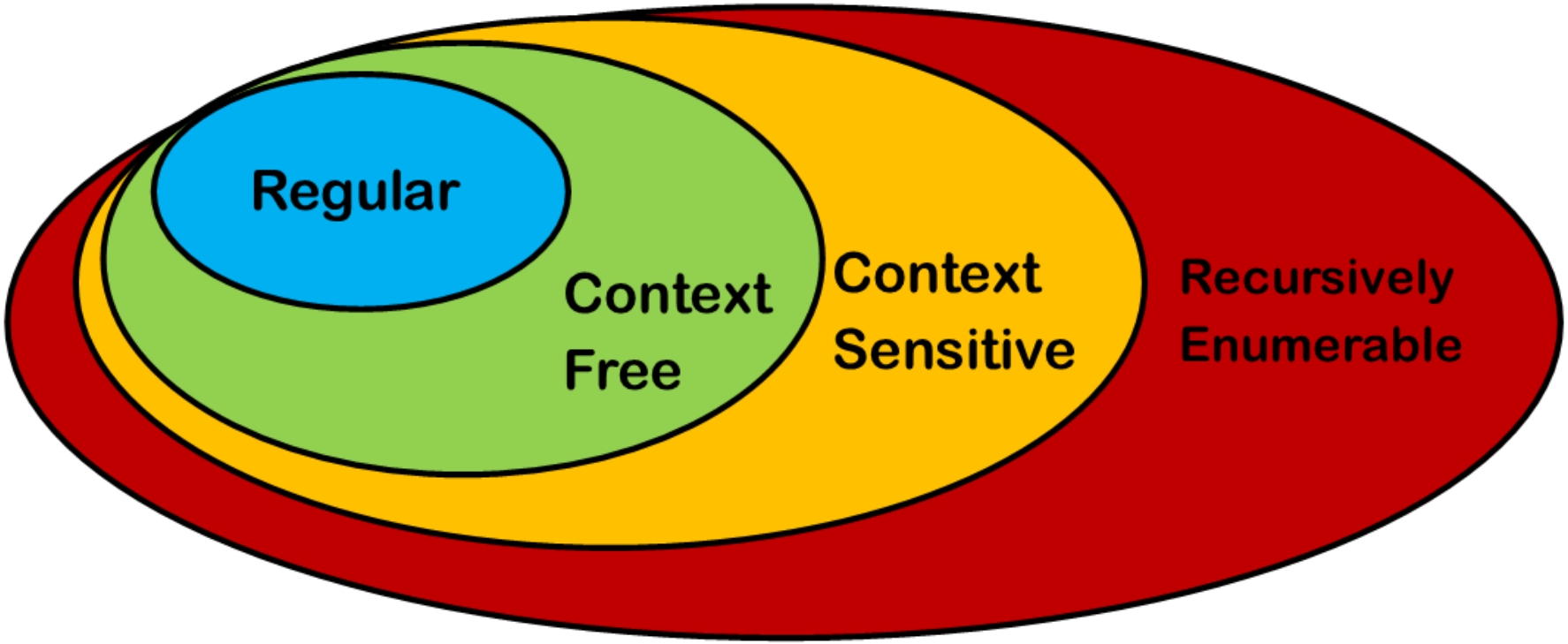
$$L_T(G) = \left\{ \begin{array}{cc} \text{S} & \text{S} \\ \text{NP} \quad \text{VP} & \text{NP} \quad \text{VP} \\ \text{N} \quad \text{N} \quad \text{Aux} \quad \text{AP} & \text{N} \quad \text{N} \quad \text{Aux} \quad \text{AP} \\ \text{human} \quad \text{language} \quad \text{isn't} \quad \text{A} & \text{language} \quad \text{human} \quad \text{isn't} \quad \text{A} \\ \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \text{subregular} & \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \text{subregular} \end{array} \right\}$$

$$L_s(G) = \{\textit{human language isn't subregular, language human isn't subregular}\}$$

- Usually called a *Context Free (string) Grammar* (CFG)

Context-Free Languages

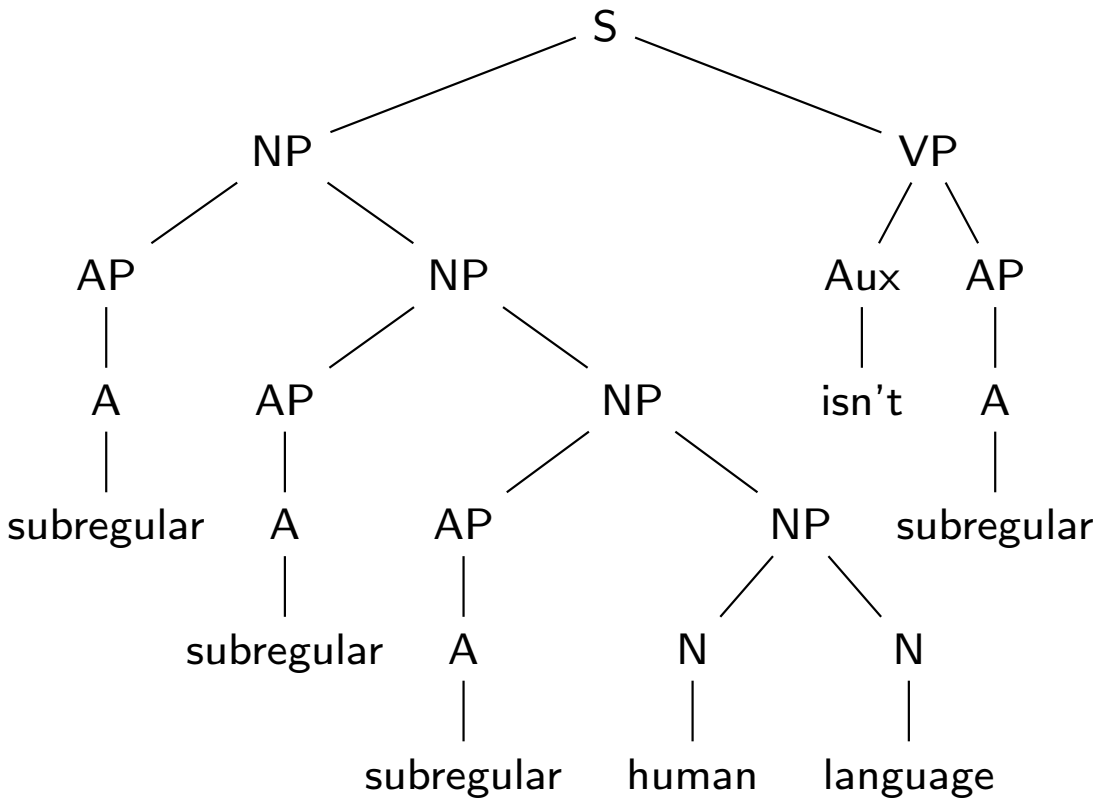
(Sub)regular over *trees*, but Context Free over *strings*.



Context Free Grammars: infinite use of finite means

If G has recursion, the language is infinite

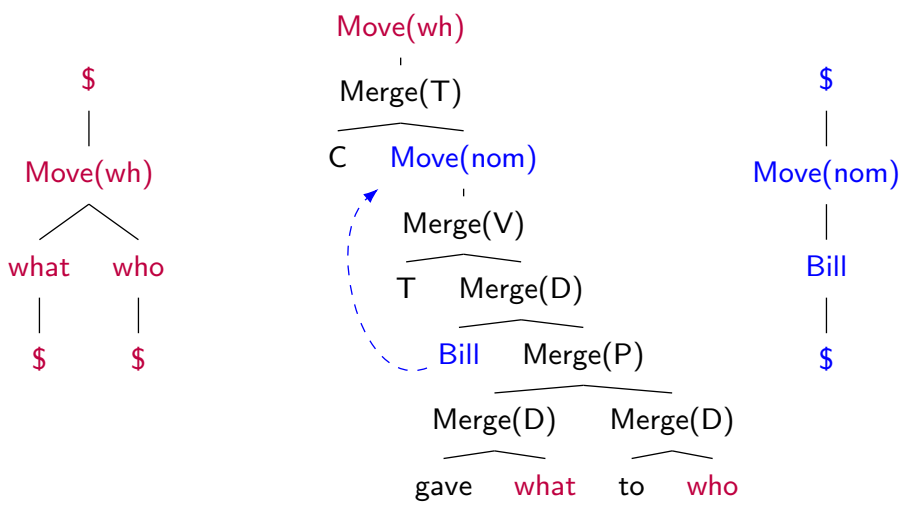
$$
\begin{aligned}
S &\rightarrow \text{NP VP} \\
NP &\rightarrow \text{N N} \\
NP &\rightarrow \text{AP NP} \\
VP &\rightarrow \text{Aux AP} \\
AP &\rightarrow \text{A} \\
N &\rightarrow \text{human} \\
N &\rightarrow \text{language} \\
Aux &\rightarrow \text{isn't} \\
A &\rightarrow \text{regular}
\end{aligned}
$$



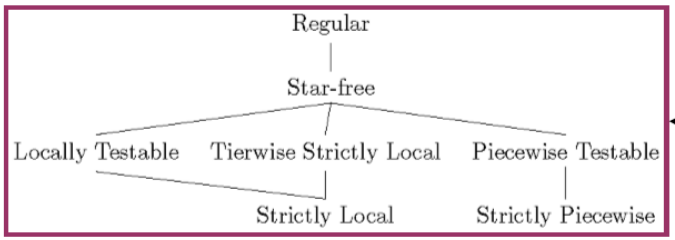
Are human languages context free?

No, but this is getting warmer! A lot of things about human language syntax is captured by Strictly Local tree languages.
Hypothesis (Graf, 2018): syntactic *derivations* are **Tierwise** Strictly Local

Blocking Many-to-One Movement

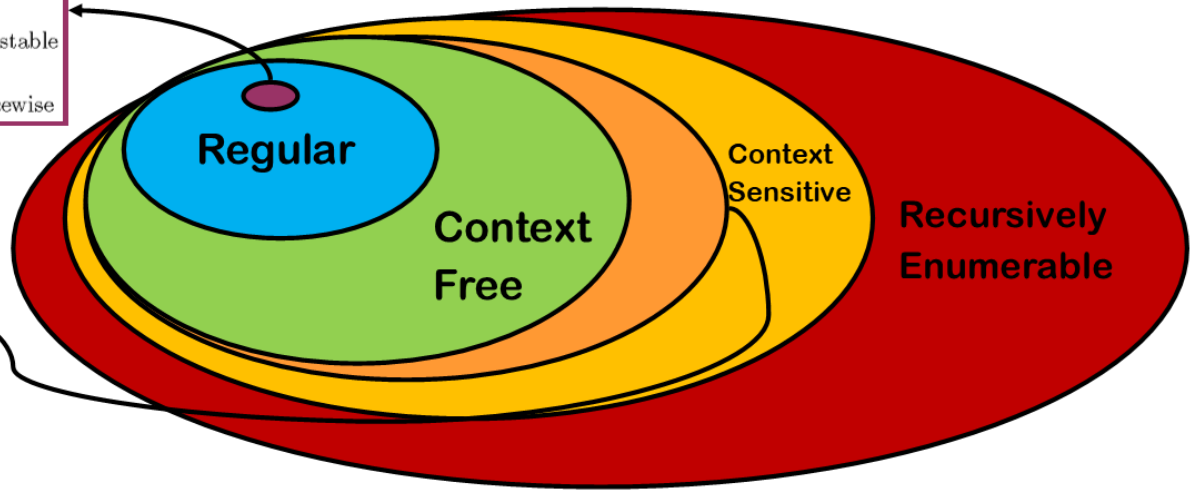


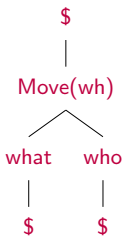
Chomsky Hierarchy

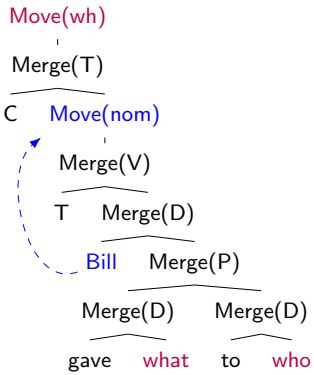


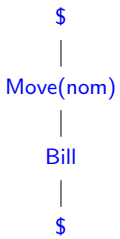
Subregular Hierarchy

Mildly Context Sensitive









Universal Grammar: math style

UG means babies know that their grammars will be Tierwise Strictly Local

Human phonotactics are Tierbased Strictly Local. Human syntactic derivations are also TSL!

Thank you! Questions?

References

Everett, Daniel Leonard. 2009. *Don't sleep, there are snakes: Life and language in the amazonian jungle*. Profile Books.

Graf, Thomas. 2018. The surprising simplicity of syntax: Derivation trees, subregular complexity, and what it implies for language and cognition. Invited talk, November 26, Integrated Language Science and Technology Seminar, University of Pennsylvania, Philadelphia, PA.

Hauser, Marc D, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *science* 298:1569–1579.

Heinz, Jeffrey, Chetan Rawal, and Herbert G Tanner. 2011. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 58–64. Association for Computational Linguistics.

Jardine, Adam, and Jeffrey Heinz. 2016. Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics* 4:87–98.