

---

# Inheritance and the CCG lexicon

Mark McConville  
University of Edinburgh

Postgraduate Conference in Linguistics at Manchester  
4 March 2005



## The Chomsky hierarchy

- type 0: recursively enumerable languages
- type 1: context-sensitive languages
- type 2: context-free languages
- type 3: regular languages

Containment hierarchy

- type 3  $\subset$  type 2, type 2  $\subset$  type 1, . . .

## Natural classes and automata

- type 0: Turing machines
- type 1: linear bounded automata
- type 2: pushdown automata
- type 3: finite state machines

Also a containment hierarchy

## Phrase structure grammars

- type 0:  $\phi \rightarrow \psi$
- type 1:  $\phi \rightarrow \psi$  where  $|\phi| < |\psi|$
- type 2:  $X \rightarrow \phi$
- type 3:  $X \rightarrow a \phi$

## Formal languages and human languages

- the human languages: a natural class of formal language
- the language faculty (“universal grammar”): a family of automata/grammars

## Human languages?

- type 0: recursively enumerable languages
- type 1: context-sensitive languages
- type 2: context-free languages
- **type 3: regular languages?**

## Human languages?

- type 0: recursively enumerable languages
- type 1: context-sensitive languages
- **type 2: context-free languages?**
- type 3: regular languages

## Recap: Formal languages and human languages

- the human languages: a natural class of formal language
- the language faculty (“universal grammar”): a family of automata/grammars
- human languages  $\subseteq$  context-free languages

BUT . . .

- context-free languages  $\not\subseteq$  human languages

## Left-gapping languages

- John loves Mary and Bill loves Kate
- John LOVES Mary and Bill loves Kate
- John \_ Mary and Bill loves Kate

## Ross' observation

- SOV languages
- left-gapping vs. non-left-gapping languages
- two combinations — both context-free
  - SOV and left-gapping — many
  - SOV and non-left-gapping — few

## Formal languages and human languages

- the human languages: a natural class of formal language
- the language faculty (“universal grammar”): a family of automata/grammars
- human languages  $\subseteq$  context-free languages
- context-free languages  $\not\subseteq$  human languages
  - SOV/left-gapping languages are “more human” than SOV/non-left-gapping languages

SO . . . context-free grammars/pushdown automata overgenerate

## A solution: Markedness

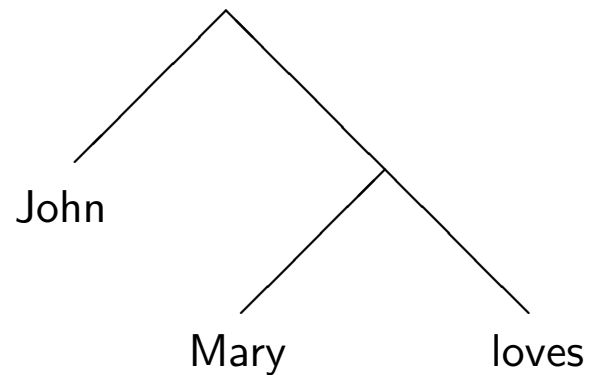
- languages are ranked
- e.g. by the “size” of the smallest grammar
- AIM: a family of grammars which ranks probable human languages more highly than improbable human languages

## CFGs and human language competence

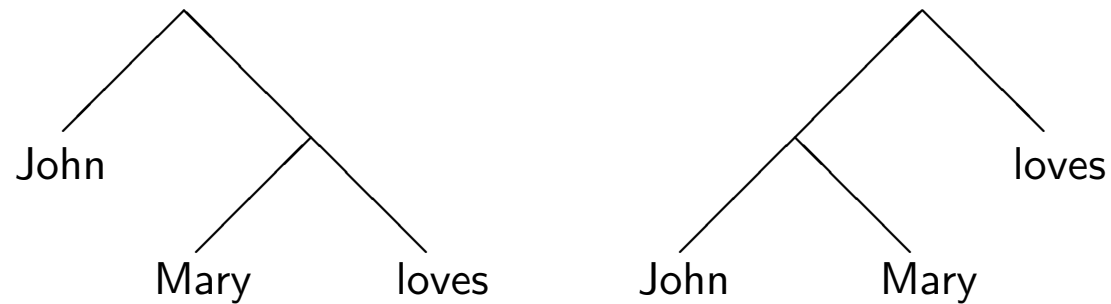
- $X \rightarrow \phi$
- the smallest CFG for an SOV/left-gapping language is no smaller than the smallest CFG for the corresponding SOV/non-left-gapping language
- the smallest CFG for an SOV/left-gapping language is *bigger* than the smallest CFG for the corresponding SOV/non-left-gapping language

## SOV/non-left-gapping

S → NP VP  
VP → NP V



## SOV/left-gapping



$S \rightarrow NP VP$   
 $VP \rightarrow NP V$   
 $S \rightarrow NC V$   
 $NC \rightarrow NP NP$

## CFGs and human language competence

- the smallest CFG for an SOV/left-gapping language is *bigger* than the smallest CFG for the corresponding SOV/non-left-gapping language
- CFG family: SOV/non-left-gapping languages are “more human” than SOV/left-gapping languages

## Alternative: Combinatory categorial grammars

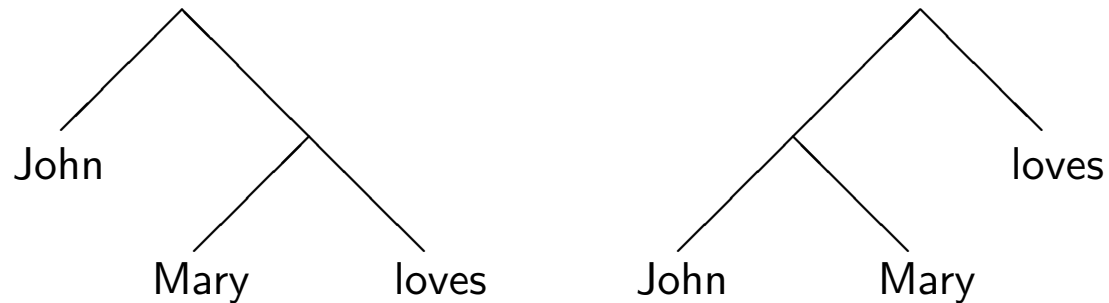
- CCGs generate all and only the context-free languages
- CCGs assume flexible constituency

# CCGs and flexible constituency

John ⊢ NP

Mary ⊢ NP

loves ⊢ S\NP\NP



## Combinatory categorial grammars

- CCGs generate all and only the context-free languages
- CCGs assume flexible constituency
- the CCG for an SOV/left-gapping language is the CCG for an SOV language
- CCG family: SOV/left-gapping languages are “more human” than SOV/non-left-gapping languages

## Summary so far

- the human languages are a natural class of formal language
- human languages  $\subseteq$  context-free languages
- context-free languages  $\not\subseteq$  human languages
- families of grammars rank the languages they generate
- AIM: family of grammars which ranks probable human languages more highly than improbable ones
- Ross' observation: flexible categorial grammars are better than traditional phrase structure grammars

## My project

- many statistical universals of human language are NOT captured by flexible categorial grammars
  - Greenbergian basic word order correlations

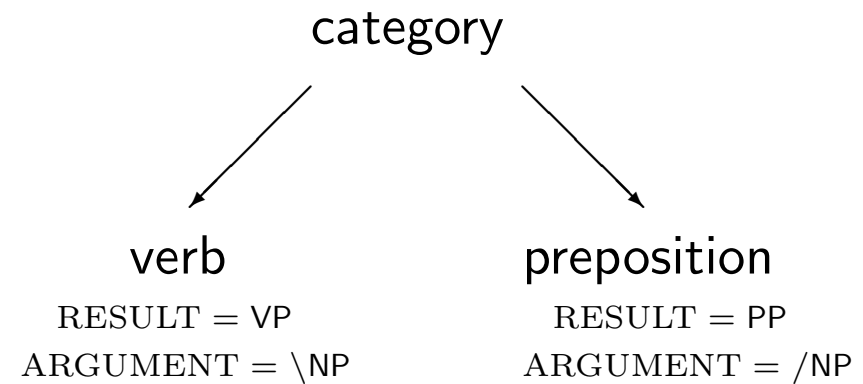
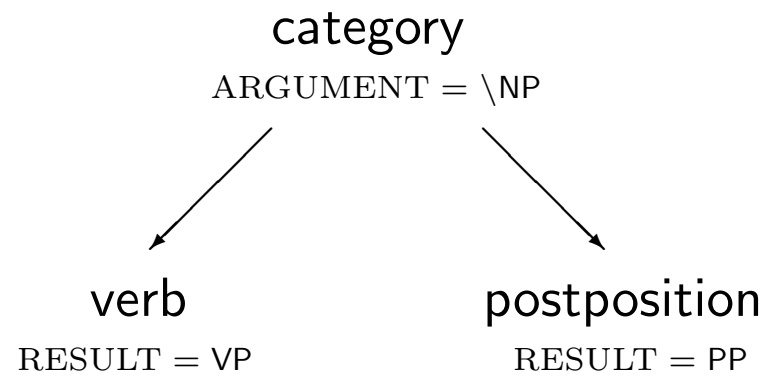
## A Greenberg universal

4. “With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional”
  - OV languages
  - OP languages vs. PO languages
  - two possibilities
    - OV/OP — many
    - OV/PO — few
  - PROBLEM: the smallest CCG of an OV/OP language is NO SMALLER than the smallest CCG of the equivalent OV/PO language

## My project

- many statistical universals of human language: NOT captured by flexible categorial grammars
  - Greenbergian basic word order correlations
- SOLUTION: organising the CCG lexicon as an inheritance hierarchy

# OV/PO languages



## My project

- many statistical universals of human language: NOT captured by flexible categorial grammars
  - Greenbergian basic word order correlations
- SOLUTION: organising the CCG lexicon as an inheritance hierarchy
- inheritance hierarchies: independently necessary for optimal encoding of lexical information
- explanatory aspects — incremental processability