

---

# CCG and Markedness

Mark McConville

ESSLLI: CCG and Linguistic Diversity  
8-12 August 2005



## Summary so far

### CCG

- is a restrictive grammar formalism
- makes clear, falsifiable predictions about the limits of linguistic diversity
- explains implicational universal  $P \cap Q = \emptyset$  if
  - there are CCGs for both  $P$  languages and  $Q$  languages
  - there is no CCG for a  $P \cap Q$  language

## The Fixed Subject Constraint

- the woman who John loves (1)
- the man who loves Mary (2)
- the woman who Bill thinks that John loves (3)
- \*the man who Bill thinks that loves Mary (4)

i.e. Subjects cannot be extracted across a complementiser.

## FSC in French

la femme que Jean aime (5)

the woman whom Jean loves

l'homme qui aime Marie (6)

the man who loves Marie

la femme que Bill croit que Jean aime (7)

the woman whom Bill believes that Jean loves

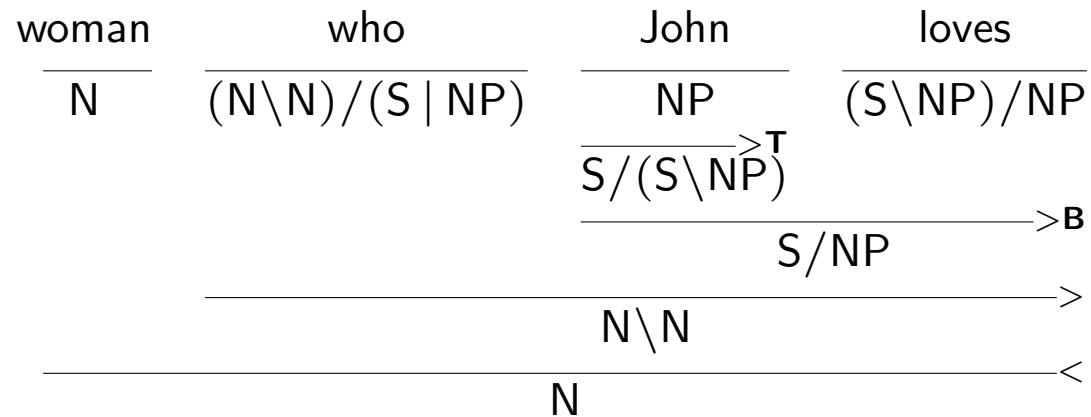
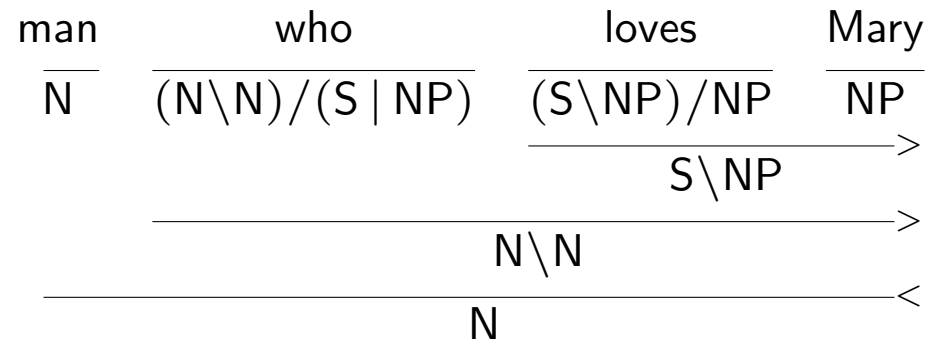
\*l'homme qui Bill croit qu'aime Marie (8)

the man who Bill thinks that loves Marie

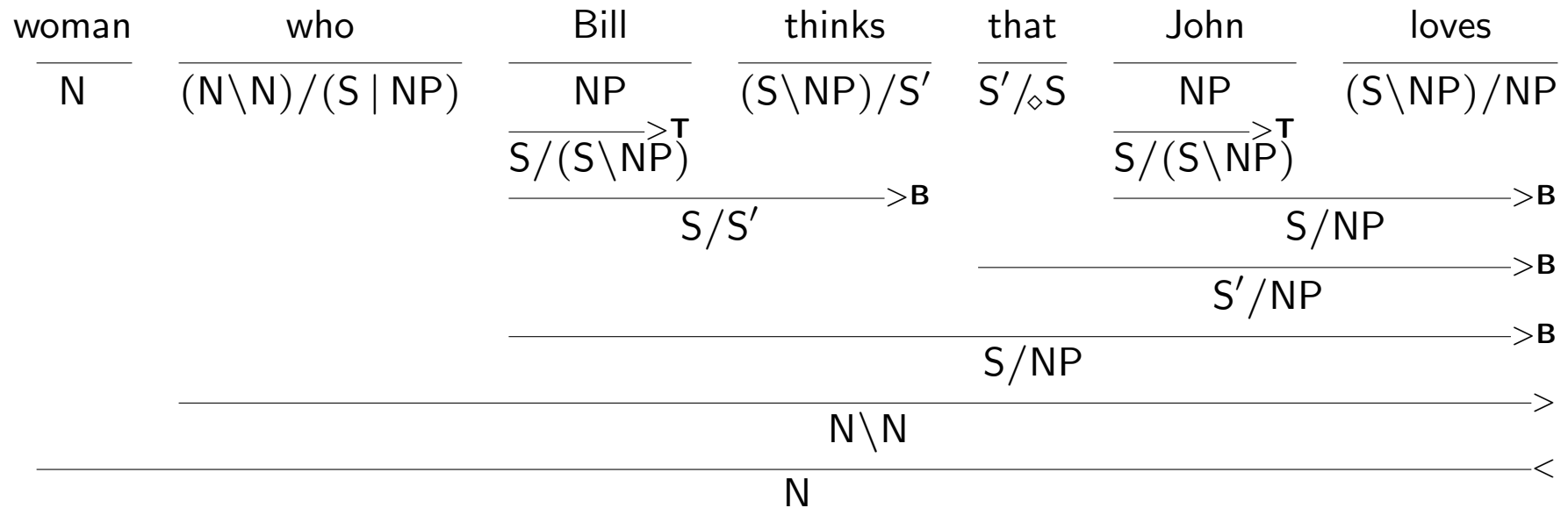
## The FSC in CCG

who  $\vdash (N \setminus N) / (S \mid NP)$   
loves  $\vdash (S \setminus NP) / NP$   
thinks  $\vdash (S \setminus NP) / S'$   
that  $\vdash S' / \diamond S$   
John, Mary, Bill  $\vdash NP$

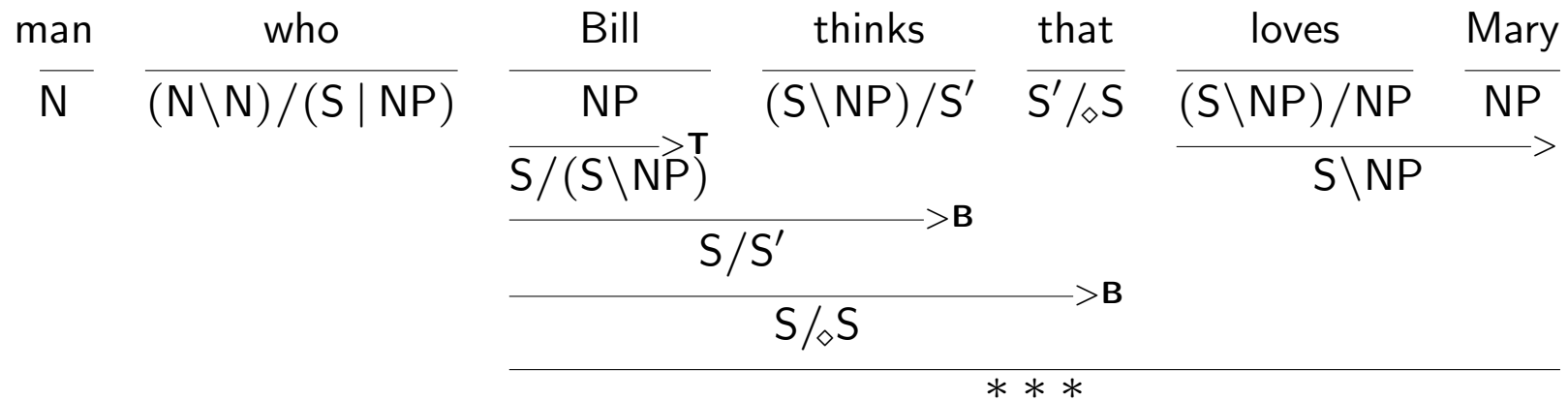
## Deriving simple English relatives



## Unbounded object relativisation



## Unbounded subject relativisation?



The FSC is predicted for English and French since:

- verbs are  $(S \setminus NP) / \$$
- complementiser is  $S' / \diamond S$

## FSC in Italian

la donna che Gianni ama (9)

the woman whom Gianni loves

l'uomo che ama Maria (10)

the man who loves Maria

la donna che Bill credo che Gianni ama (11)

the woman whom Bill believes that Gianni loves

l'uomo che Bill credo che ama Maria (12)

the man who Bill believes that loves Maria

i.e. in Italian, subjects *can* be extracted across a complementiser



## CCG for Italian

che  $\vdash (N \setminus N) / (S / NP)$

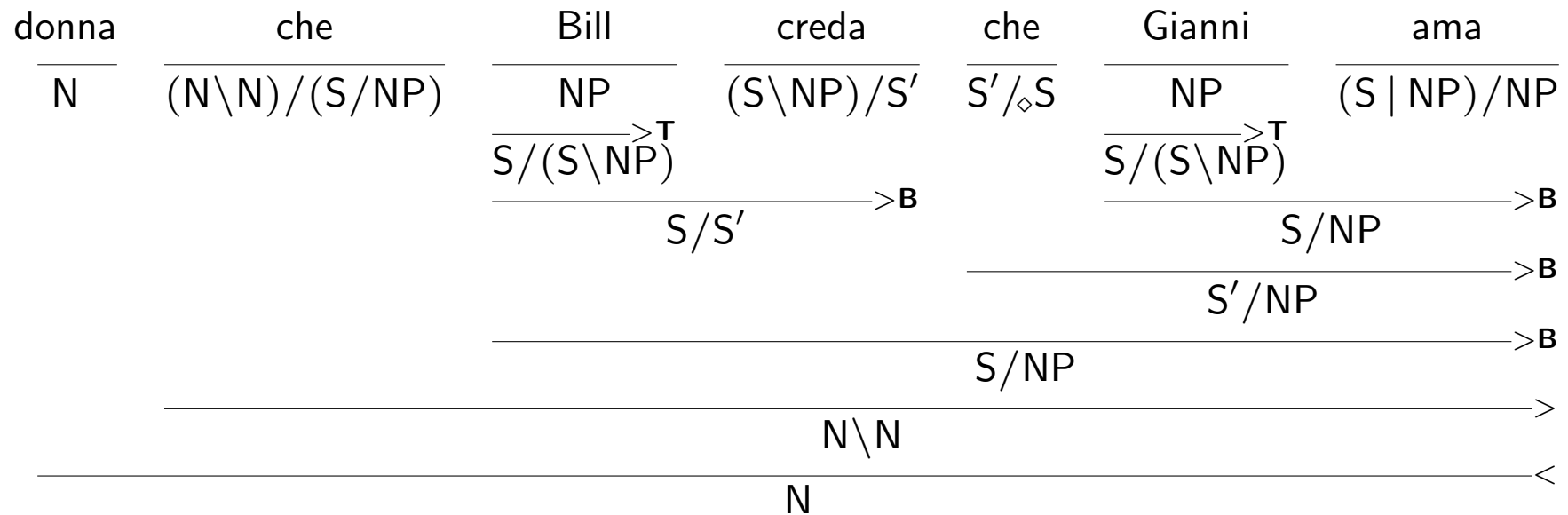
ama  $\vdash (S \mid NP) / NP$

creda  $\vdash (S \setminus NP) / S'$

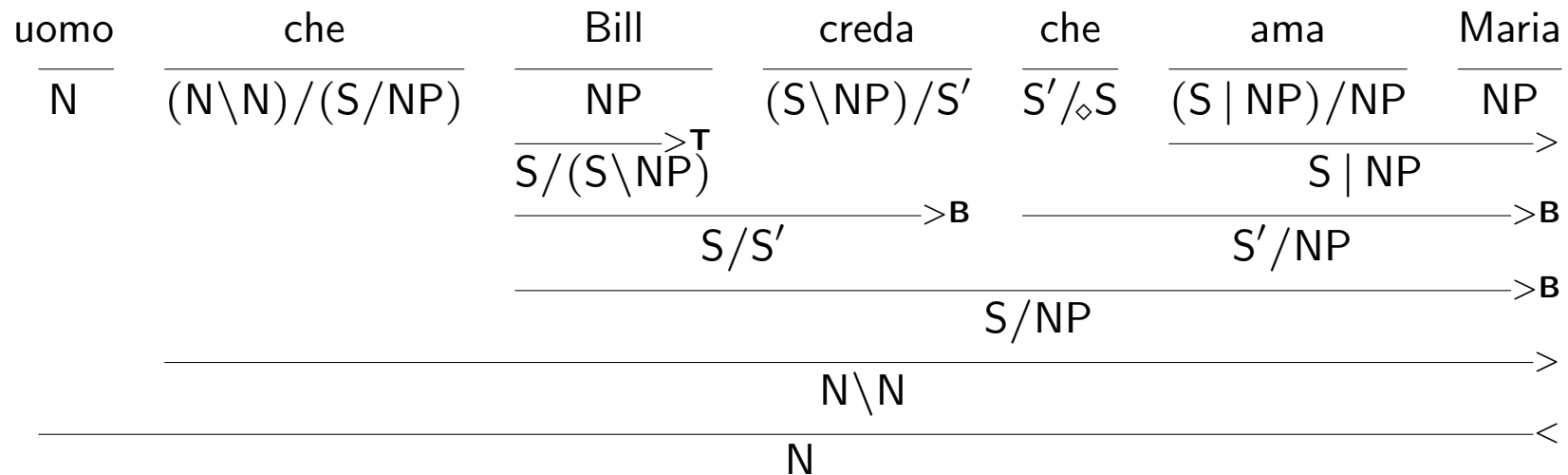
che  $\vdash S' / \diamond S$

Gianni, Maria, Bill  $\vdash NP$

## Unbounded object relatives in Italian



## Unbounded subject relatives in Italian



## The Fixed Subject Constraint in CCG

CCG predicts that English and French will satisfy the FSC, since

- verbs are  $(S \setminus NP) / \$$
- complementisers are  $S' / \diamond S$

CCG predicts that Italian will violate the FSC, since

- verbs are  $(S \mid NP) / \$$
- complementisers are  $S' / \diamond S$

## FSC as implicational universal

1. Domain: Romance languages (including English)
2. Typology:
  - FSC vs. FSC'
  - FI vs. FI'
3. Observations:
  - $FSC \cap FI$
  - $FSC \cap FI'$  e.g. English, French
  - $FSC' \cap FI$  e.g. Italian, Spanish, Romanian, Middle French
  - $FSC' \cap FI'$

## Explaining implicational universals in CCG

Formalism  $F$  explains implicational universal  $P \cap Q = \emptyset$  if

- there are  $F$ -grammars for  $P$  languages
- there are  $F$ -grammars for  $Q$  languages
- there is no  $F$ -grammar for a  $P \cap Q$  language

Does CCG explain the *pro*-drop parameter in Romance?

$$FSC \cap FI = \emptyset$$

$$FSC' \cap FI' = \emptyset$$

## CCG for an $FSC \cap FI'$ language

who  $\vdash (N \setminus N) / (S \mid NP)$   
loves  $\vdash (S \setminus NP) / NP$   
thinks  $\vdash (S \setminus NP) / S'$   
that  $\vdash S' / \diamond S$   
John, Mary, Bill  $\vdash NP$

## CCG for an $FSC' \cap FI$ language

che  $\vdash (N \setminus N) / (S / NP)$

ama  $\vdash (S \mid NP) / NP$

creda  $\vdash (S \setminus NP) / S'$

che  $\vdash S' / \diamond S$

Gianni, Maria, Bill  $\vdash NP$

## CCG for an $FSC \cap FI$ language?

che  $\vdash (N \setminus N) / (S \mid NP)$

ama  $\vdash (S \setminus NP) / NP$

creda  $\vdash (S \setminus NP) / S'$

che  $\vdash S' / \diamond S$

Gianni, Maria, Bill  $\vdash NP$

Gianni, Maria, Bill  $\vdash S \setminus (S \setminus NP)$

i.e. subject extraction blocked

## CCG for an $FSC' \cap FI'$ language

che  $\vdash (N \setminus N) / (S / NP_{ant})$

che  $\vdash (N \setminus N) / (S / NP_{\theta})$

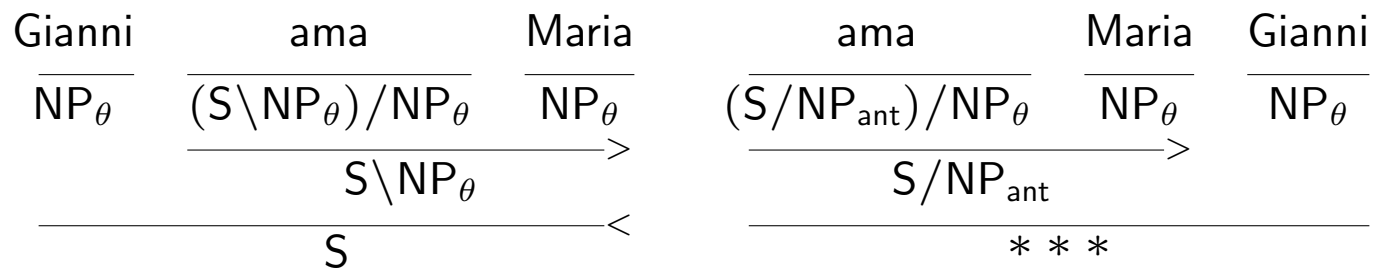
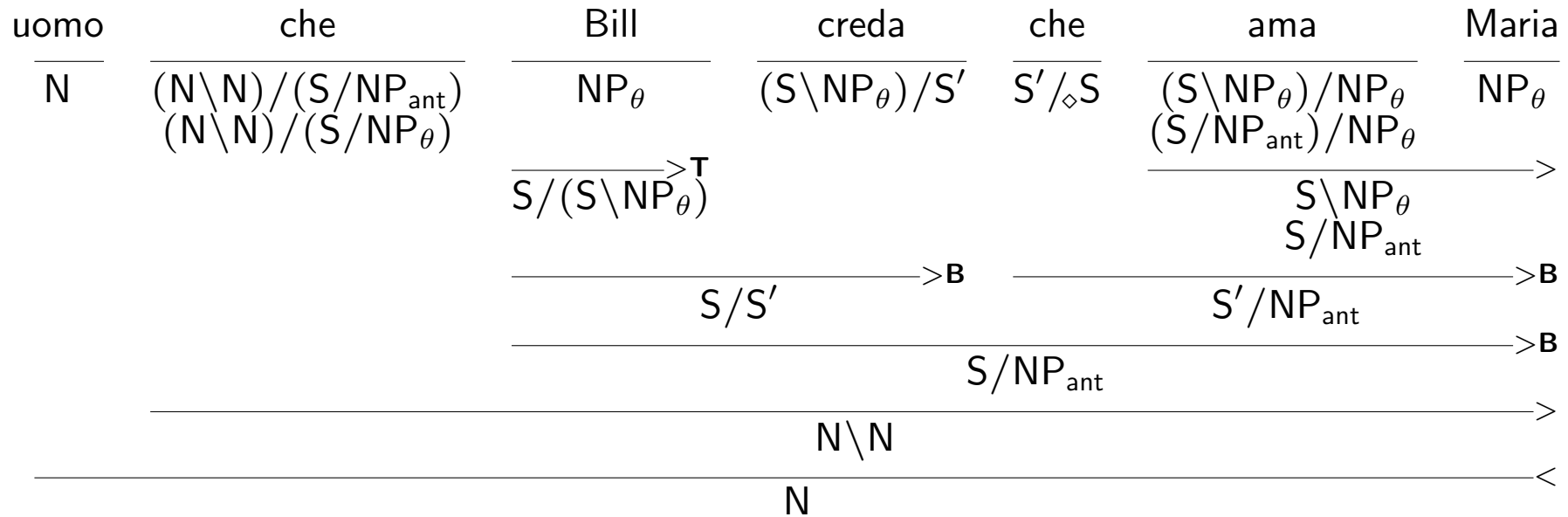
ama  $\vdash (S \setminus NP_{\theta}) / NP_{\theta}$

ama  $\vdash (S / NP_{ant}) / NP_{\theta}$

creda  $\vdash (S \setminus NP_{\theta}) / S'$

che  $\vdash S' /_{\diamond} S$

Gianni, Maria, Bill  $\vdash NP_{\theta}$



## Does CCG explain the *pro-drop* parameter?

Formalism  $F$  explains implicational universal  $P \cap Q = \emptyset$  if

- there is no  $F$ -grammar for a language in  $P \cap Q$ , and . . .

But: CCGs exist for  $FSC \cap FI$  and  $FSC' \cap FI'$  languages.

Important: the *simplest* CCGs for  $FSC \cap FI$  and  $FSC' \cap FI'$  languages are *more complex* than CCGs of their  $FSC \cap FI'$  and  $FSC' \cap FI$  equivalents.

## Explanation in CCG

Formalism  $F$  explains implicational universal  $P \cap Q = \emptyset$  if

- there are  $F$ -grammars for  $P$  languages
- there are  $F$ -grammars for  $Q$  languages
- the simplest  $F$ -grammars for  $P \cap Q$  languages are more complex than  $F$ -grammars of their  $P \cap Q'$  and  $P' \cap Q$  equivalents

Therefore: CCG explanations of some implicational universals invoke considerations of the comparative simplicity (i.e. the 'size') of two grammars.

## The problem of the CCG lexicon — part I

The generative capacity of CCG is the MCS languages.

But: not every MCS language over the words of English is an equally likely natural language.

Two approaches:

1. substantive constraints on CCGs
2. evaluation metric for CCGs
  - a measure of grammar quality

## How to rank CCGs?

Aim: to define an evaluation metric on the class of CCGs which reflects the comparative likelihood of the languages they generate being natural languages.

At least two possibilities:

1. by size
2. by processability

## Ranking CCGs by size

CCG  $G_1 = \langle A_1, S_1, L_1 \rangle$  is better than CCG  $G_2 = \langle A_2, S_2, L_2 \rangle$  iff:

- $G_1$  is at least as communicative as  $G_2$
- $L_1$  is smaller than  $L_2$

## Ranking CCGs by processability

CCG  $G_1 = \langle A_1, S_1, L_1 \rangle$  is better than CCG  $G_2 = \langle A_2, S_2, L_2 \rangle$  iff:

- $G_1$  is at least as communicative as  $G_2$
- the sentences generated by  $G_1$  are on average parsed more easily than those generated by  $G_2$

cf. Jack Hawkins, Mark Johnson, Glynn Morrill

## The problem of the CCG lexicon — part II

CCG lexicons for natural languages are highly redundant.

John		John
he	love-s	me
the girl		you
girl-s		him
I		us
you	love	them
we		the girl
they		girl-s
girl-s		girl-s

## A redundant CCG

$\langle \text{John}, \text{NP}_x \rangle$	$\langle \text{John}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{girl}, \text{N}_{\text{sg}} \rangle$	$\langle \text{s}, \text{N}_{\text{pl}} \setminus \text{N}_{\text{sg}} \rangle$
$\langle \text{s}, \text{NP}_{\text{subj}} \setminus \text{N}_{\text{sg}} \rangle$	$\langle \text{s}, \text{NP}_{\text{obj}} \setminus \text{N}_{\text{sg}} \rangle$
$\langle \text{the}, \text{NP}_x / \text{N}_{\text{sg}} \rangle$	$\langle \text{the}, \text{NP}_{\text{obj}} / \text{N}_{\text{sg}} \rangle$
$\langle \text{the}, \text{NP}_{\text{subj}} / \text{N}_{\text{pl}} \rangle$	$\langle \text{the}, \text{NP}_{\text{obj}} / \text{N}_{\text{pl}} \rangle$
$\langle \text{I}, \text{NP}_{\text{subj}} \rangle$	$\langle \text{me}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{we}, \text{NP}_{\text{subj}} \rangle$	$\langle \text{us}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{you}, \text{NP}_{\text{subj}} \rangle$	$\langle \text{you}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{he}, \text{NP}_x \rangle$	$\langle \text{him}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{they}, \text{NP}_{\text{subj}} \rangle$	$\langle \text{them}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{love}, (\text{S} \setminus \text{NP}_{\text{subj}}) / \text{NP}_{\text{obj}} \rangle$	$\langle \text{s}, ((\text{S} \setminus \text{NP}_x) / \text{NP}_{\text{obj}}) \setminus ((\text{S} \setminus \text{NP}_{\text{subj}}) / \text{NP}_{\text{obj}}) \rangle$

## Functionality and atomicity

**Criterion of functionality** the ideal lexicon is a function from morphemes to category labels

**Criterion of atomicity** the ideal lexicon is a mapping onto a set of atomic category labels

What do we need to add the CCG to allow NL grammars which satisfy these criteria? How do we retain CCG-ness? How to avoid reinventing HPSG?

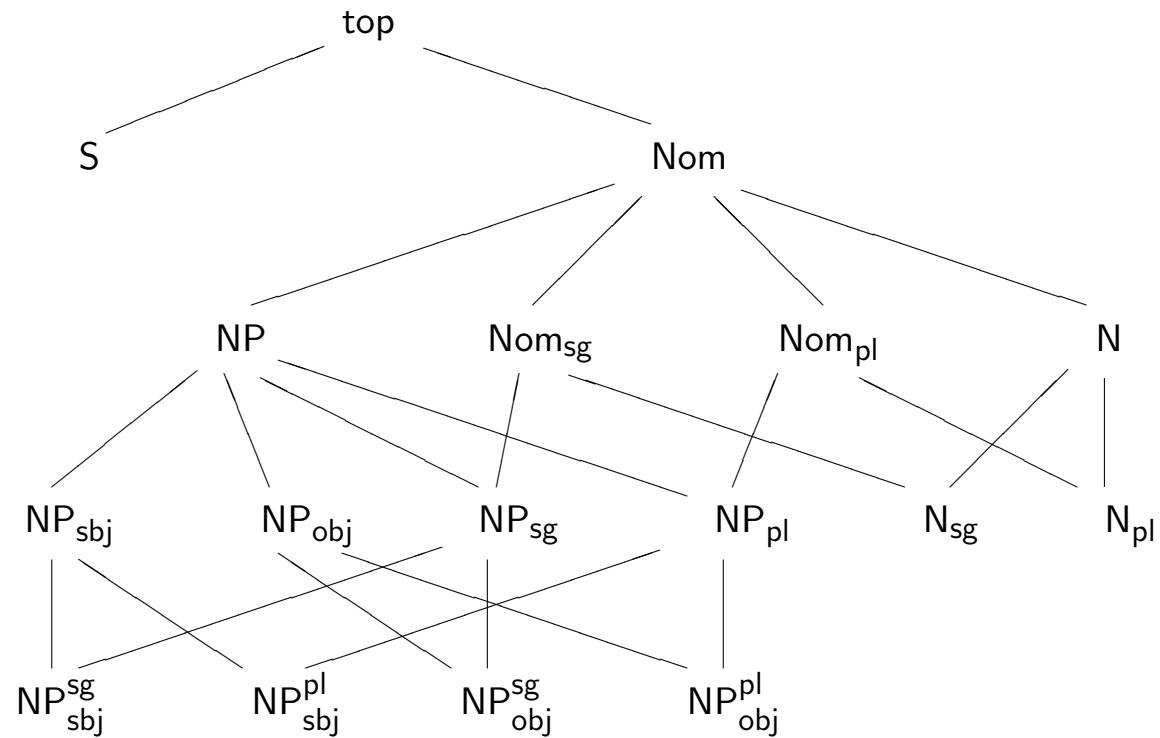
e.g. macros/families, set-based categories, feature structures, type hierarchies, constraint inheritance, lexical rules?

## Adding type hierarchies to CCG

A CCG over alphabet  $\Sigma$  is an ordered 4-tuple  $\langle A, \sqsubseteq, S, L \rangle$  where:

1.  $\langle A, \sqsubseteq \rangle$  is a type hierarchy of saturated category symbols
2.  $S$  is a distinguished subset of  $A$
3.  $L$  is a finite mapping from  $\Sigma$  to CCG categories over  $A$

# A type hierarchy



## A less redundant CCG lexicon

$\langle \text{John}, \text{NP}_{\text{sg}} \rangle$	$\langle \text{girl}, \text{N}_{\text{sg}} \rangle$
$\langle \text{I}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$	$\langle \text{me}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{we}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$	$\langle \text{us}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{you}, \text{NP}_{\text{pl}} \rangle$	$\langle \text{he}, \text{NP}_{\text{sbj}}^{\text{sg}} \rangle$
$\langle \text{him}, \text{NP}_{\text{obj}} \rangle$	$\langle \text{they}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$
$\langle \text{them}, \text{NP}_{\text{obj}} \rangle$	$\langle \text{love}, (\text{S} \setminus \text{NP}_{\text{sbj}}^{\text{pl}}) / \text{NP}_{\text{obj}} \rangle$
$\langle \text{s}, \text{Nom}_{\text{pl}} \setminus \text{N}_{\text{sg}} \rangle$	$\langle \text{the}, \text{NP}_{\text{sg}} / \text{N}_{\text{sg}} \rangle$
$\langle \text{the}, \text{NP}_{\text{pl}} / \text{N}_{\text{pl}} \rangle$	$\langle \text{s}, ((\text{S} \setminus \text{NP}_{\text{sbj}}^{\text{sg}}) / \text{NP}_{\text{obj}}) \setminus ((\text{S} \setminus \text{NP}_{\text{pl}}) / \text{NP}) \rangle$

## Inheritance hierarchies of lexical category labels

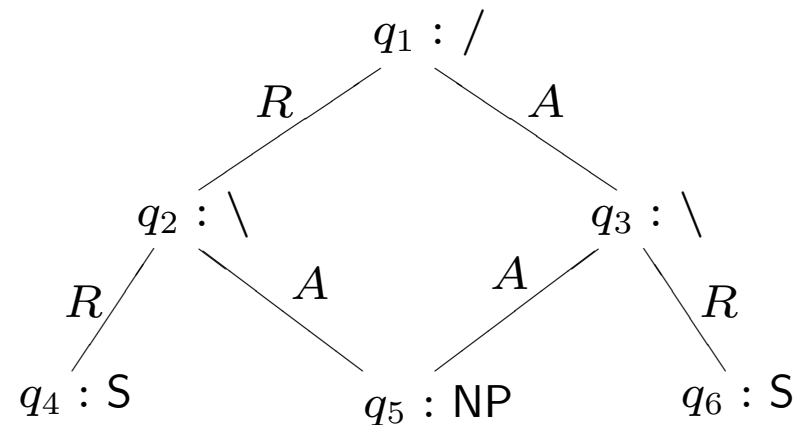
- category models vs. category descriptions
- satisfaction relation
- CCG lexicon is a mapping from morphemes to sets of category descriptions
  - indirect relation between a forms and its categories

## Inheritance-based CCG

A CCG over alphabet  $\Sigma$  is an ordered 4-tuple  $\langle A, \sqsubseteq_A, B, \sqsubseteq_B, S, L \rangle$  where:

1.  $\langle A, \sqsubseteq_A \rangle$  is a type hierarchy of saturated category symbols
2.  $\langle B, \sqsubseteq_B \rangle$  is an inheritance hierarchy over the set of category descriptions over  $A$
3.  $S$  is a distinguished subset of  $A$
4.  $L$  is a function from  $\Sigma$  to  $B \cup A$

## Category models



A category model over alphabet  $A$  of saturated category symbols is an ordered 5-tuple  $\langle Q, Res, Arg, V_S, V_A \rangle$ , where . . .

## Category descriptions

The set of category descriptions over alphabet  $A$  of saturated category symbols is defined as the smallest set  $\Phi$  such that:

1.  $A \subseteq \Phi$
2. for all  $\delta \in \{/, \backslash\}$ , (SLASH  $\delta$ )  $\in \Phi$
3. for all  $\mu \in \{\star, \diamond, \times\}$ , (MOD  $\mu$ )  $\in \Phi$
4. for all  $\phi \in \Phi$ , (ARG  $\phi$ )  $\in \Phi$
5. for all  $\phi \in \Phi$ , (RES  $\phi$ )  $\in \Phi$

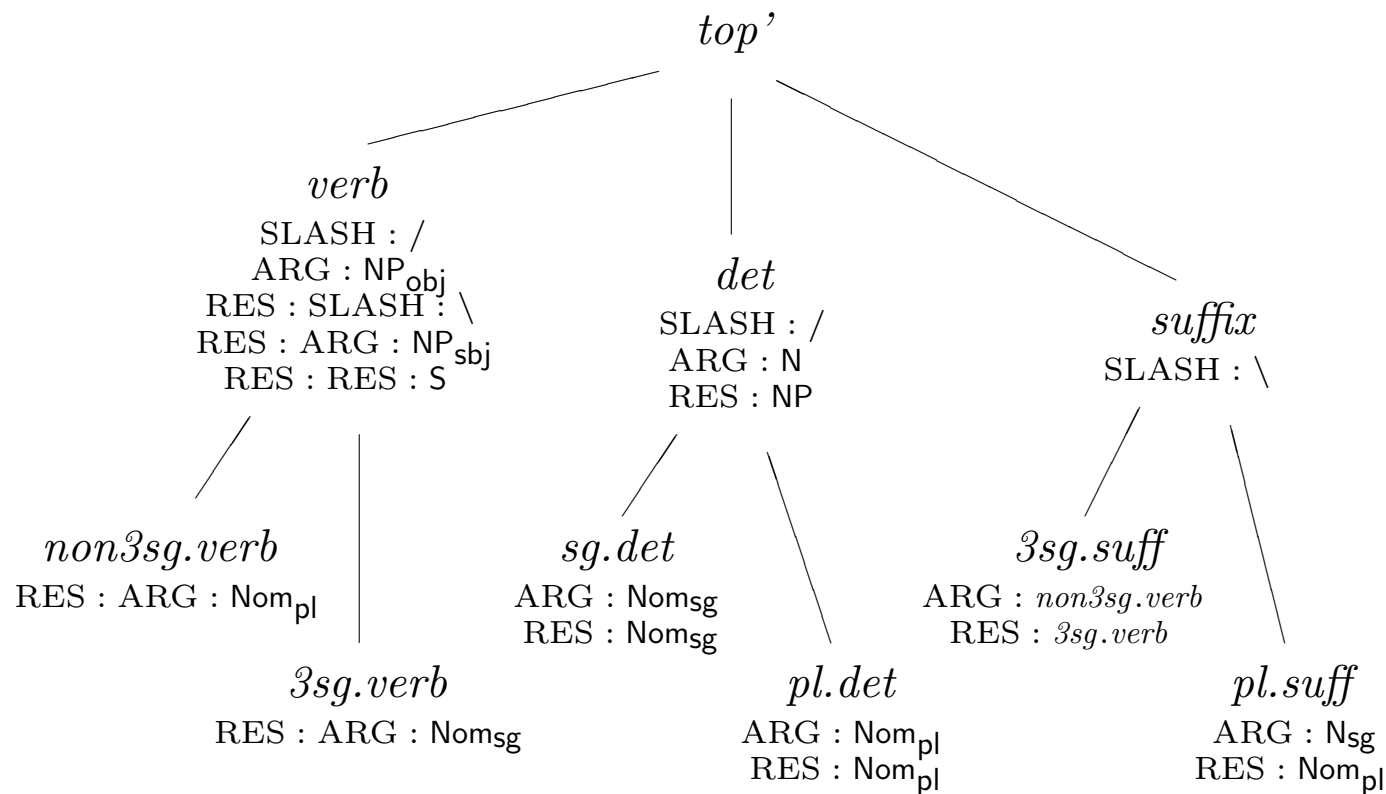
## Satisfaction

Category model  $\mathcal{S}$  over type hierarchy  $\langle A, \sqsubseteq \rangle$  satisfies category description  $\phi$  over  $A$ , written  $\mathcal{S} \models \phi$ , if and only if  $\mathcal{S}, q \models \phi$ , where  $q$  is the root of  $\mathcal{S}$ .

Category model  $\mathcal{S} = \langle Q, Res, Arg, V_S, V_M, V_A \rangle$  over type hierarchy  $\langle A, \sqsubseteq \rangle$  locally satisfies category description  $\phi$  over  $A$ , at point  $q \in Q$ , written  $\mathcal{S}, q \models \phi$ , if and only if:

1. where  $\phi \in A$ : for some  $\psi \in A$  such that  $\phi \sqsubseteq \psi$ ,  $V_A(q) = \psi$
2. where  $\phi = (\text{SLASH } \delta)$ :  $V_S(q) = \delta$
3. where  $\phi = (\text{MOD } \mu)$ :  $V_M(q) = \mu$
4. where  $\phi = (\text{ARG } \psi)$ :  $\mathcal{S}, Arg(q) \models \psi$
5. where  $\phi = (\text{RES } \psi)$ :  $\mathcal{S}, Res(q) \models \psi$

# A lexical inheritance hierarchy



## A non-redundant CCG

$\langle \text{John}, \text{NP}_{\text{sg}} \rangle$	$\langle \text{girl}, \text{N}_{\text{sg}} \rangle$
$\langle \text{I}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$	$\langle \text{me}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{we}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$	$\langle \text{us}, \text{NP}_{\text{obj}} \rangle$
$\langle \text{you}, \text{NP}_{\text{pl}} \rangle$	$\langle \text{he}, \text{NP}_{\text{sbj}}^{\text{sg}} \rangle$
$\langle \text{him}, \text{NP}_{\text{obj}} \rangle$	$\langle \text{they}, \text{NP}_{\text{sbj}}^{\text{pl}} \rangle$
$\langle \text{them}, \text{NP}_{\text{obj}} \rangle$	$\langle \text{love}, \text{non3sg.verb} \rangle$
$\langle \text{s}, \text{suffix} \rangle$	$\langle \text{the}, \text{det} \rangle$

Note: both functional and atomic.

## Advantages of Inheritance-based CCG

- allows functional, atomic lexicons for NLs
- lexical projection is still finite
- not HPSG i.e. clear distinction between formal and substantive universals
- can be applied to Greenbergian universals
- but: don't forget processability