
Towards a Performance Theory of Markedness in Combinatory Categorical Grammar

Mark McConville

University of Edinburgh
21 April 2005



What is CCG?

- a grammar formalism i.e. a theory of a class of formal languages
- categories
 - saturated e.g. S, NP
 - unsaturated e.g. $S \backslash NP$, $(S \backslash NP) / NP$
- grammars

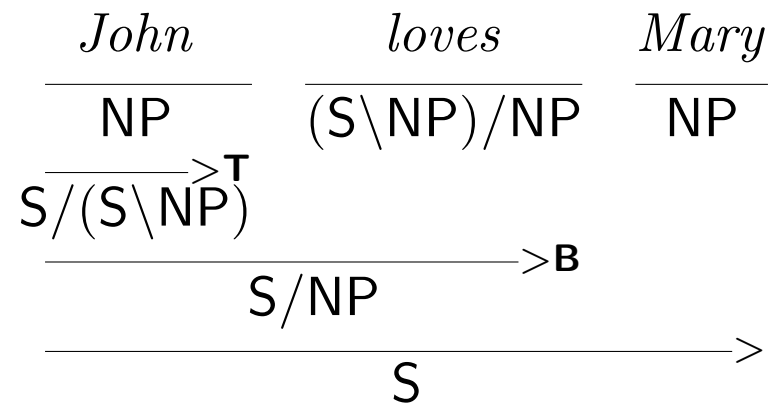
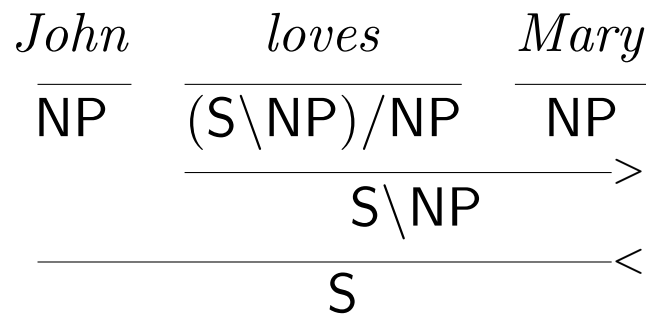
John \vdash NP

Mary \vdash NP

loves $\vdash (S \backslash NP) / NP$

What is CCG?

- universal combinatory operations
 - application i.e. $X/Y \ Y \Rightarrow X$
 - composition, raising, cross composition, substitution
- derivations



Why do I like CCG?

- CCG is simple, formally explicit, computable, semantically transparent
- CCG is linguistically motivated
 - unbounded dependencies
 - coordination
- CCG is mildly context sensitive (MCS)
 - cross serial dependencies — YES
 - doubly unbounded long distance scrambling — NO
- CCG is psycholinguistically plausible — incremental processing

Markedness in CCG

- every natural language is a CCG language
- but: not every CCG language is an equally probable natural language
 - e.g. VO prepositional vs. VO postpositional
- a theory of markedness — a ranking of the CCG languages corresponding to typological data
- *intrinsic* markedness — a function of grammar size
 - the smaller the grammar, the less marked the language

Inheritance-based CCG

- saturated categories are organised into a type hierarchy
- lexical categories are organised into an inheritance hierarchy over a flexible constraint language
- non-redundant CCG lexicons for natural languages
- an intrinsic markedness ordering which predicts a number of well-known statistical universals

Hawkins' performance theory of markedness

- inspired by the same problem
- context-free grammar (CFG)
- but: not every CF language is an equally probable natural language
- an *extrinsic* measure of markedness
 - languages are ranked according to processability
 - natural CF languages are more easily parsed than unnatural ones
- processability modelled by “early immediate constituency” (EIC)

How does EIC work?

- every node in a tree has a mother node constructor (MNC)
 - the leaf node from which its existence and nature can be deduced
- every non-leaf node in a tree has a constituent recognition domain (CRD)
 - the sequence of leaf nodes from the MNC of its leftmost child to the MNC of its rightmost child
- every non-leaf node in a tree has an EIC value
 - the left-to-right IC-to-word ratio of its CRD
 - a real number between 0 (worst) and 1 (optimal)

How does EIC work?

- every tree has an EIC value
 - the average of the EIC values of its non-leaf nodes
 - again, a real number between 0 (worst) and 1 (optimal)
- every CFG has an EIC value
 - the average of the EIC values of the trees it generates
- so: CFGs are ranked according to their EIC values
 - ‘natural’ CFGs have higher EIC values than ‘unnatural’ CFGs

Predictions of Hawkins' theory

- consistent headedness — uniform right- or left-branching
- head-initial languages with optional rightward extraction are less marked than those without (i.e. heavy NP shift, extraposition)
- head-final languages with optional leftward extraction are less marked than those without (i.e. *S'* preposing in Japanese)
- light elements (e.g. pronouns) tend to the left
- heavy elements (e.g. relative clauses) tend to the right

EIC and markedness in CCG?

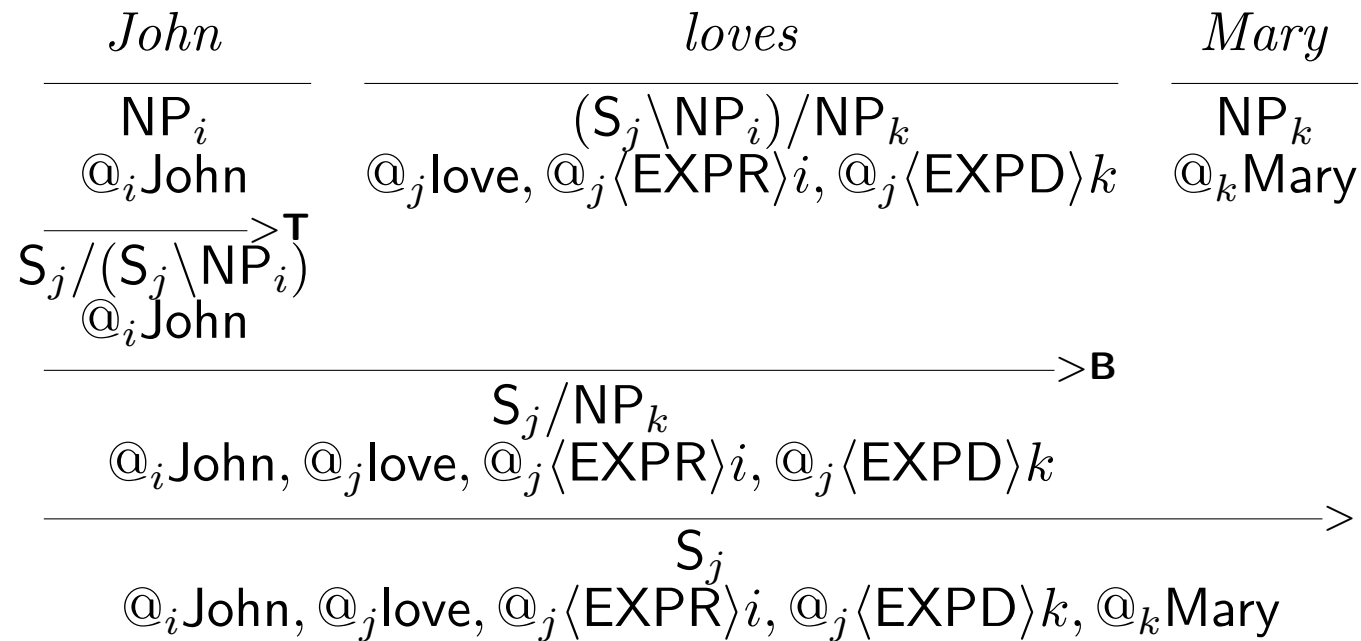
- parsing with CCGs does not involve building syntactic structures
- parsing builds semantic representations incrementally
 - (flattened) hybrid logic dependency terms

John \vdash $\text{NP}_i : @_i \text{John}$

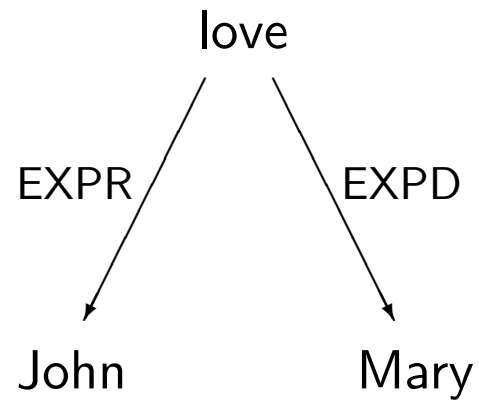
Mary \vdash $\text{NP}_i : @_i \text{Mary}$

loves \vdash $(S_i \setminus \text{NP}_j) / \text{NP}_k : @_i \text{love}, @_i \langle \text{EXPR} \rangle_j, @_i \langle \text{EXPD} \rangle_k$

A derivation



$@_i\text{John}, @_j\text{love}, @_j\langle\text{EXPR}\rangle_i, @_j\langle\text{EXPD}\rangle_k, @_k\text{Mary}$



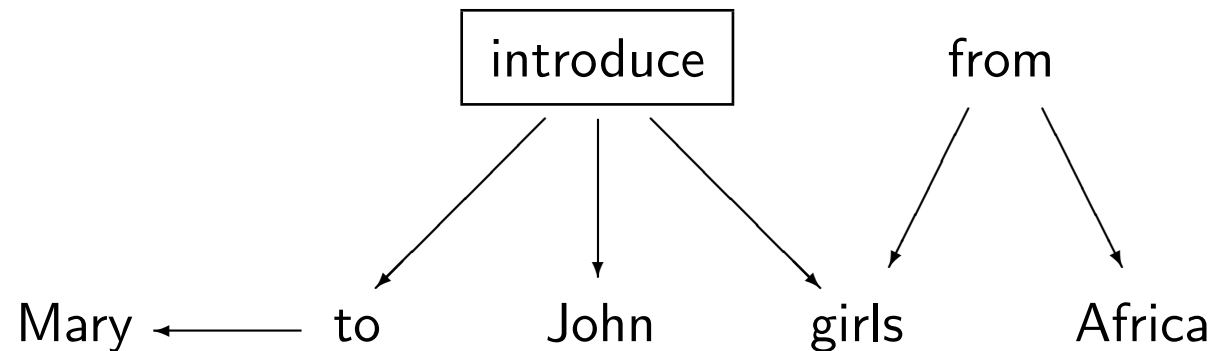
EIC and markedness in CCG?

- EIC values for normal form CCG derivations?
- EIC values for the models underlying HLD terms?

EIC for HLD models

- every node in the HLD model of a sentence has a MNC
 - the word which contributes the node's semantic content

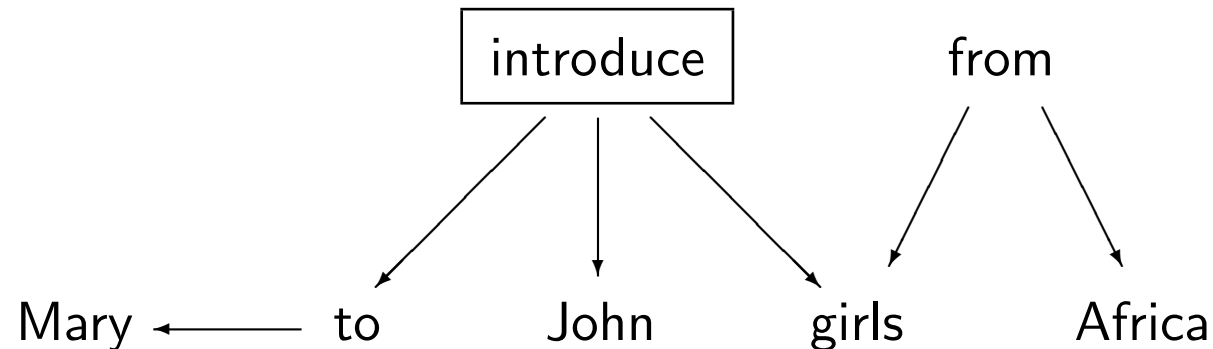
e.g. *John* *introduced* *girls from Africa to Mary*



EIC for HLD models

- every node in the HLD model of a sentence has a RD
 - the shortest sequence of words including its own MNC and those of all its dependents

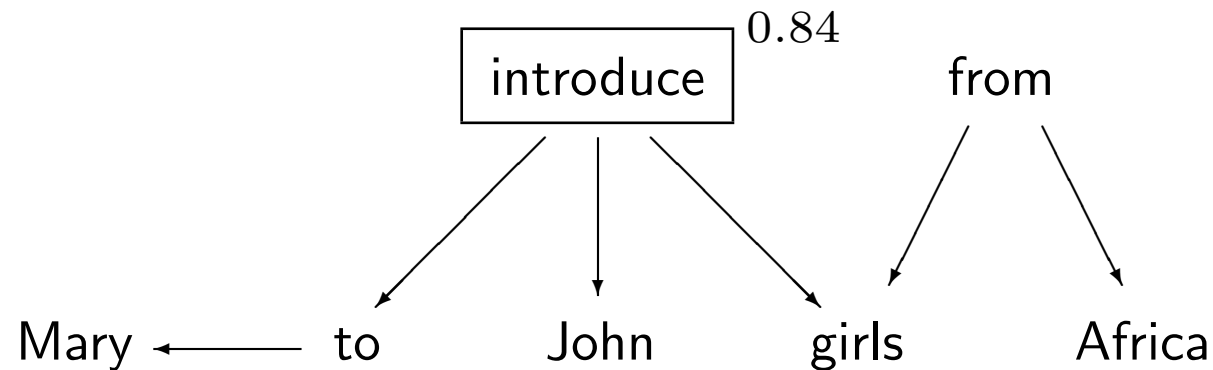
e.g. *John introduced girls from Africa to* *Mary*



EIC for HLD models

- every node in the HLD model of a sentence has an EIC value
 - the left-to-right dependent-to-word ratio of its CRD

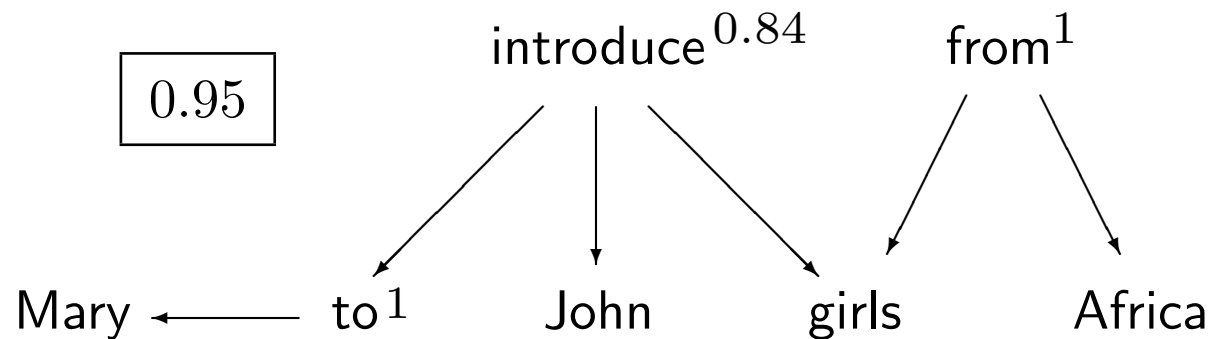
e.g. *John introduced girls from Africa to* *Mary*



EIC for HLD models

- every HLD model of a sentence has an EIC value
 - the average of the EIC values of its non-end nodes

e.g. *John introduced girls from Africa to Mary*



EIC for HLD models

- every CCG has an EIC value
 - the average of the best EIC values of the HLD models for every sentence generated by the grammar
- every CCG language has an EIC value
 - the best EIC value from all the CCGs which generate it
- a performance-based ranking of CCG languages

Conclusions

- simulate Hawkins' EIC metric as a performance-based markedness ranking for CCG
 - on semantic representations rather than syntactic structures
- explains all the same statistical universals as Hawkins
 - headedness
 - reordering transformations
 - “heaviness” phenomena

Questions

- implications for parsing strategies?
 - as many HLD terms as early as possible?
 - rate of building local dependency structures?
- Greenberg's universal 41 (i.e. case \approx SOV)?
 - early access to who is doing what to whom?
- some kinds of dependents more important than others?
- proof nets and incremental processing?