The Interface of Syntax and Information Structure

Steedman's CCG approach

Detmar Meurers
(based on joint preparation with Kordula De Kuthy)
LING795K, OSU, Spring 2005

A CCG approach to information structure

Combinatory Categorial Grammar (CCG; Steedman 2000a,b)

- CCG in a nutshell
- Structure, intonation, and information structure
- The two dimensions of information structure
- Combinatory Prosody

CCG in a nutshell

- Syntactically potent elements such as verbs are associated with a syntactic category that identifies them as *functions* and specifies the type and directionality of their arguments and the type of their result.
- A "result leftmost" notation is used:
  - α/β is a rightward-combining functor over a domain β into a range α
  - $\alpha\backslash\beta$ is the corresponding leftward-combining functor.
  - α and β may themselves be functional categories.

  (1) proved := $(S\backslash NP)/NP$

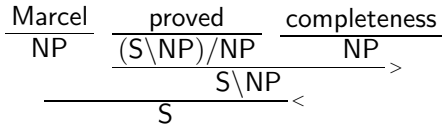
Rules and derivations

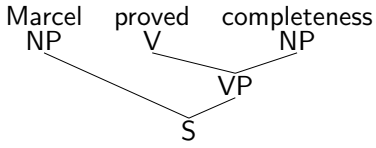
- Functor categories can combine with their arguments by the following rules:

  (2) Forward application (>)
  $X/Y \quad Y \quad \Rightarrow \quad X$

  (3) Backward application (<)
  $Y \quad X\backslash Y \quad \Rightarrow \quad X$

- Derivations are written as shown below, on the left side. Note the direct correspondence to the upside-down constituency tree shown on the right.





Semantics and Principle of Type Transparency

- The lexical categories can be augmented with an explicit identification of their semantic interpretation and the rules of functional application are accordingly expanded with an explicit semantics.

  (4) proved $:= (S\backslash NP)/NP : prove'$

  (5) Forward application ($>$)
  $X/Y : f \quad Y : a \quad \Rightarrow \quad X : fa$

- The semantic interpretation of all combinatory rules is fully determined by the *Principle of Type Transparency*:

  All syntactic categories reflect the semantic type of the associated logical form, and all syntactic combinatory rules are type-transparent versions of one of a small number of semantic operations over functions including application, composition, and type-raising.

Example derivation with semantics

$$\dfrac{\dfrac{\text{Marcel}}{\text{NP} : \text{marcel}'} \quad \dfrac{\dfrac{\text{proved}}{(\text{S}\backslash\text{NP})/\text{NP} : \text{prove}'} \quad \dfrac{\text{completeness}}{\text{NP} : \text{completeness}'}}{\text{S}\backslash\text{NP} : \text{prove}'\ \text{completeness}'}\,{>}}{\text{S} : \text{prove}'\ \text{completeness}'\ \text{marcel}'}\,{<}$$

More rule schemata

CCG includes linguistically motivated rule schemata such as the one for coordination of constituents of like type shown below:

(6) Coordination ($< \& >$)
$X\ conj\ X \quad \Rightarrow \quad X$

Combinators

- In order to account for coordination of contiguous strings that do not constitute traditional constituents, CCG allows certain operations on functions called "combinators", including the rule of functional composition in (7).

  (7) Forward composition ($>$**B**)
  $X/Y : f \quad Y/Z : g \quad \Rightarrow \quad X/Z : \lambda x.f(gx)$

- CCG includes type-raising rules, which turn arguments into functions over functions-over-such-arguments.

  (8) Forward type-raising ($>$**T**)
  $X : a \quad \Rightarrow \quad T/(T\backslash X) : \lambda f.fa$

  (9) Backward type-raising ($<$**T**)
  $X : a \quad \Rightarrow \quad T\backslash(T/X) : \lambda f.fa$

  X ranges over argument categories (e.g., NP and PP). The rules are order-preserving, e.g., (8) can turn an NP into a rightward-looking function over leftward functions, preserving the linear order of subjects and predicates.

Non-standard surface structures

- Complement-taking verbs like *think*, VP/S, can compose with fragments like *Marcel proved*, S/NP, which accounts for right-node raising (10), and also provides the basis for an analysis of unbounded dependencies (11).

  (10) [I disproved]$_{S/NP}$ and [you think that Marcel proved]$_{S/NP}$ completeness.

  (11) the result that [you think that Marcel proved]$_{S/NP}$

  Strings such as *you think that Marcel proved* are taken to be surface constituents of type S/NP.

Non-standard surface structures are licensed throughout

- Steedman assumes that the non-traditional constituents motivated for right-node raising and similar coordinations are also possible constituents of non-coordinate sentences like *Marcel proved completeness*.

$$\dfrac{\dfrac{\dfrac{\text{Marcel}}{\text{NP}:\text{marcel}'}}{\text{S}/(\text{S}\backslash\text{NP}):\lambda\text{f.f marcel}'}\,{>}\mathbf{T} \qquad \dfrac{\text{proved}}{(\text{S}\backslash\text{NP})/\text{NP}:\text{prove}'}}{\text{S}/\text{NP}:\lambda\text{x.prove}'\,\text{x marcel}'}\,{>}\mathbf{B} \qquad \dfrac{\dfrac{\text{completeness}}{\text{NP}:\text{completeness}'}}{\text{S}\backslash(\text{S}/\text{NP}):\lambda\text{p.p completeness}'}\,{<}\mathbf{T}}{\text{S}:\text{prove}'\,\text{completeness}'\,\text{marcel}'}\,{<}$$

$$\dfrac{\dfrac{\dfrac{\text{Marcel}}{\text{NP}:\text{marcel}'}}{\text{S}/(\text{S}\backslash\text{NP}):\lambda\text{f.f marcel}'}\,{>}\mathbf{T} \qquad \dfrac{\text{proved}}{(\text{S}\backslash\text{NP})/\text{NP}:\text{prove}'} \qquad \dfrac{\dfrac{\text{completeness}}{\text{NP}:\text{completeness}'}\,{>}\mathbf{B}}{(\text{S}\backslash\text{NP})\backslash((\text{S}\backslash\text{NP})/\text{NP}):\lambda\text{p.p completeness}'}\,{<}\mathbf{T}}{\text{S}\backslash\text{NP}:\lambda\text{y.prove}'\,\text{completeness}'\,\text{y}}\,{<}}{\text{S}:\text{prove}'\,\text{completeness}'\,\text{marcel}'}\,{<}$$

- The Principle of Type Transparency guarantees that all such non-standard derivations yield identical interpretations.

Motivating non-standard surface structures

- According to Steedman (2000a), the non-standard surface structures are not spurious ambiguities but relevant since they subsume the intonation structures needed to explain the possible intonation contours for sentences of English.
- Intonational boundaries contribute to determining which of the possible combinatory derivations is intended.
- The interpretations of the constituents that arise from these derivations are related to semantic distinctions of information structure and discourse focus.
- Steedman's claims:
  - Where intonational boundaries are present, they contribute to disambiguation.
  - Conversely, any such boundaries must be consistent with *some* syntactic derivation, or ill-formedness will result.

Examples for impossible intonation boundaries

(12) a. * (Three mathematicians) (in ten derive a lemma).
  b. * (Seymour prefers the nuts) (and bolts approach).
  c. * (They only asked whether I knew the woman who chaired) (the zoning board).

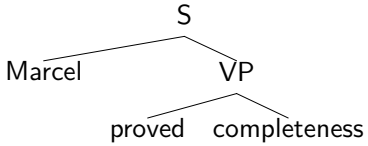
Syntactic structure and intonation

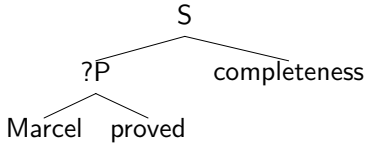
Steedman's claims:

- *Surface structure* and *information structure* coincide, the latter simply consisting in the interpretation associated with a constituent analysis of the sentence.
- *Intonation* coincides with *surface structure*, and hence information structure, in the sense that all intonational boundaries coincide with syntactic boundaries (but not all syntactic boundaries are intonationally marked).

As a result, fragments such as *Marcel proved* in (13c), are not only prosodic constituents but surface syntactic constituents, complete with interpretations.

(13) a. Marcel proved completeness.

b.


c.


Intonation and Information Structure

- A sequence of one or more pitch accents followed by a boundary is referred to as an *intonational phrasal tune*.
- *Claim:* phrasal tunes in this sense are associated with specific discourse meanings distinguishing information type and/or propositional attitude.

(14) Q: I know who proved soundness. But who proved COMPLETENESS?
A: (MARCEL) (proved COMPLETENESS).
H* L L+H* LH%

(15) Q: I know which result Marcel PREDICTED. But which result did Marcel PROVE?
A: (Marcel PROVED) (COMPLETENESS).
L+H* LH% H* LL%

- *Evidence:* Exchanging the answer tunes between the two contexts in (14) and (15) yields complete incoherence.

The two dimensions of Information Structure

- *Theme* and *Rheme*:
  - The *theme* is to be thought of as that part of an utterance which connects it to the rest of the discourse.
  - The *rheme* is that part of an utterance that advances the discussion by contributing novel information.

  (Note that Steedman's Theme/Rheme corresponds to Background/Focus in the terminology otherwise used in this course.)
- *Focus* and *Background*:
  - The information marked by the pitch accent is called the *focus*, distinguishing theme focus and rheme focus, where necessary.
  - The term *background* is used for the part unmarked by pitch accent or boundary.

  (Note that Steedman uses Focus in a narrow, phonological sense.)

Theme and Rheme and their intonational realization

Steedman observes the following relationship for English:

- The L+H* LH% tune is associated with the *theme*.
- The H* L and H* LL% tunes (among others) are associated with the *rheme*.

Intonationally unmarked themes/rhemes

- There also are *intonationally unmarked themes*:

(16) Q: Which result did Marcel prove?
A: (Marcel proved) (COMPLETENESS).
H* LL%

(17) Q: What do you know about Marcel?
A: (Marcel) (proved COMPLETENESS).
H* LL%

- The same contour can also occur with an *all-rheme* utterance:

(18) Guess what? (Marcel proved COMPLETENESS!)
H* LL%

Semantic characterization of theme and rheme

- Following Jackendoff (1972), the *theme* is characterized semantically via functional abstraction, using the notation of λ-calculus, as in (19), corresponding to the theme of (15) and (16).

(19) $\lambda x.prove'\ x\ marcel'$

- When such a function is supplied with an argument in the form of the rheme, it reduces to give a proposition, with the same predicate-argument relation as the canonical sentence.

(20) $prove'\ completeness'\ marcel'$

Semantic characterization of theme and rheme (cont.)

- The λ-abstraction operator is closely related to the existential quantifier $\exists$

(21) $\exists x.prove'\ x\ marcel'$

- The *theme* can be associated with the *rheme alternative set*: the set of propositions that could instantiate the corresponding existentially quantified proposition.

(22) $\left\{ \begin{array}{l} prove'\ decidability'\ marcel' \\ prove'\ soundness'\ marcel' \\ prove'\ completeness'\ marcel' \end{array} \right\}$

- The theme tune and the rheme tune can be specified in semantic terms:

(23) Theme tunes *presuppose* a rheme alternative set.
Rheme tunes *restrict* the rheme alternative set.

Focus and Background

- Within both theme and rheme, those words that contribute to distinguishing the theme and the rheme of an utterance from other alternatives made available by the context may be marked via a pitch accent.

(24) Q: I know that Marcel likes the man who wrote the musical. But who does he ADMIRE?
A: (Marcel ADMIRES) (the woman who DIRECTED the musical).
L+H* LH% H* LL%
background *focus* *background* *focus* *background*
theme *rheme*

Themes, pitch accents, and the theme alternative set

- The significance of the presence or absence of primary pitch accents within a theme lies in the prior existence of a theme differing in its translation only in those elements corresponding to the accented items.
- The presence of pitch accents in the translation of themes is marked by distinguishing the corresponding constant with an asterisk.

  (25) $\exists x.\, *admires'\, x\ marcel'$

- The set of alternative themes is called the *theme alternative set*.

  (26) $\left\{ \begin{array}{l} \exists x.admires'\, x\ marcel' \\ \exists x.likes\ marcel' \end{array} \right\}$

- Such an utterance is only felicitous if a compatible prior theme can be retrieved or accommodated (i.e., the theme alternative set contains more than one element).

Combinatory Prosody: Pitch Accents

- Six pitch accents are distinguished as markers either of theme (θ) or rheme (ρ).

  (27) θ-markers: L+H*, L*+H
  ρ-markers: H*, L*, H*+L,H+L*

- Pitch accents affect both the syntactic category and the interpretation of the words they occur on.
  - With basic types, such as NP, the effect of a θ- or ρ-marking accent is to associate with the category a value of θ or ρ on a feature INFORMATION, which is notated as NP_θ or NP_ρ.
  - With function types, such as $S\backslash NP$, the effect of a θ- or ρ-marking accent is to θ- or ρ-mark the domain and range of the function, as in $S_\rho\backslash NP_\rho$.
  - Any argument that combines with such a marked function has to be compatible with its theme- or rheme-hood.

Combinatory Prosody: Pitch Accents (cont.)

- θ- and ρ-marking happens pre-syntactically, at the level of lexical categories.

  (28) proved:= $(S_\rho\backslash NP_\rho)/NP_\rho$: $\lambda x.\lambda y.\, *prove'\, xy$
  H*

- All lexical items in a sentence are associated with a pitch accent or with the "null tone", a phonological category corresponding to the absence of any tone.
- This null tone
  - marks a syntactic category with a null information feature value η,
  - which is a variable unique to each particular occurrence of the null tone, that ranges over the theme and rheme markers θ and ρ (and nothing else except η itself).

  (29) proved:= $(S_\eta\backslash NP_\eta)/NP_\eta$: $\lambda x.\lambda y.prove'\, xy$

Combinatory Prosody: Spreading of theme and rheme

- The phonologically augmented categories allow intonational tunes to be spread over arbitrarily large constituents.

(30)

Marcel	PROVED	COMPLETENESS
	L+H*	LH%

$$\frac{\mathrm{S/(S\backslash NP)} : \lambda \mathrm{p.p\ marcel'} \qquad \mathrm{(S_\theta\backslash NP_\theta)/NP_\theta} : \lambda \mathrm{x.}\lambda \mathrm{y.}\ *\mathrm{prove'\ x\ y}}{\mathrm{S_\theta/NP_\theta} : \lambda \mathrm{x.}\ *\mathrm{prove'\ x\ marcel'}}\ >\mathrm{B}$$

Combinatory Prosody: Spreading of theme and rheme (cont.)

- Iterated compositions of the same kind have the effect of allowing the theme and rheme markers associated with the pitch accents to spread unboundedly across any sequence that forms a grammatical constituent according to the combinatory grammar.

(31) ALICE says he proved COMPLETENESS

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\text{L+H*}}{S_\theta/(S_\theta\backslash NP_\theta)} \quad \dfrac{}{(S\backslash NP)/S}}{S_\theta/S_\theta}\,{>}\mathbf{B} \quad \dfrac{}{S/(S\backslash NP)}}{S_\theta/(S_\theta\backslash NP_\theta)}\,{>}\mathbf{B} \quad \dfrac{}{(S\backslash NP)/NP}}{S_\theta/NP_\theta}\,{>}\mathbf{B}}{\text{LH\%}}$$

Combinatory Prosody: The Boundaries

- The distinction between intermediate phrases and intonational phrases:
  - *Intermediate phrases* consist of one or more pitch accents, followed by either the L or the H boundary, also known as the *phrasal tone*.
  - An *intonational phrase* consists of one or more intermediate phrases followed by an L% of H% boundary tone.
- The intermediate phrase boundaries are assigned a category which transfers the theme/rheme marking to the corresponding semantic functions θ' and ρ' via the variable η':

  (32) L, H $:= S\$_\iota\backslash S\$_\eta : \lambda f.\eta' f$ (with $S\$_\eta = S_\eta$ or mapping into S_η)

  Syntactically, it maps θ and ρ-marked categories onto identically ι-marked categories, where ι will no longer unify with η, θ or ρ. This prevents further combination with anything except similarly complete prosodic phrases.

Combinatory Prosody: The Boundaries (cont.)

- The intonational phrase boundary tones L% and H% are assigned the categories in (33). Intermediate phrase boundaries are mapped into intonational phrase boundaries.

  (33) L% $:= (S\$_\phi\backslash S\$_\eta)\backslash(S\$_\iota\backslash S\$_\eta) : \lambda f.\lambda g.[S](fg)$
  H% $:= (S\$_\phi\backslash S\$_\eta)\backslash(S\$_\iota\backslash S\$_\eta) : \lambda f.\lambda g.[H](fg)$

- Just like ι for intermediate phrases, ϕ prevents further combination with anything except similarly complete prosodic phrases.
- The modal operators $[S]$ and $[H]$ are intended to distinguish speaker's and hearer's knowledge (in a further to be worked out manner).

Two examples from Steedman (2000a)

(67)

$$\dfrac{\dfrac{\dfrac{\text{Marcel}}{S/(S\backslash NP) : \lambda p.p\ marcel'} \quad \dfrac{\overset{\text{PROVED}}{\text{L+H*}}}{(S_\theta\backslash NP_\theta)/NP_\theta : \lambda x.\lambda y.{*}prove' xy}}{S_\theta/NP_\theta : \lambda x.{*}prove' x\ marcel'}\,{>}\mathbf{B} \quad \dfrac{\dfrac{\text{L}}{S\$_\iota\backslash S\$_\eta : \lambda f.\eta' f} \quad \dfrac{\text{H\%}}{(S\$_\phi\backslash S\$_\eta)\backslash(S\$_\iota\backslash S\$_\eta) : \lambda f.\lambda g.[H](fg)}}{S\$_\phi\backslash S\$_\eta : \lambda f.[H](\eta' f)}\,{<}}{S_\phi/NP_\phi : [H](\theta'(\lambda x.{*}prove' x\ marcel'))}\,{<}$$

$\overset{\text{COMPLETENESS}}{\text{H*}}$ L L%

(68)

$$\dfrac{\dfrac{\dfrac{\text{Marcel} \quad \overset{\text{PROVED}}{\text{L+H*}} \quad \text{L} \quad \text{H\%}}{S_\phi/NP_\phi : [H](\theta'(\lambda x.{*}prove' x\ marcel'))} \quad \dfrac{\overset{\text{COMPLETENESS}}{\text{H*}}}{S_\rho\backslash(S_\rho/NP_\rho) : \lambda p.p\ {*}completeness'} \quad \dfrac{\dfrac{\text{L}}{S\$_\iota\backslash S\$_\eta : \lambda f.\eta' f} \quad \dfrac{\text{L\%}}{(S\$_\phi\backslash S\$_\eta)\backslash(S\$_\iota\backslash S\$_\eta) : \lambda f.\lambda g.[S](fg)}}{S\$_\phi\backslash S\$_\eta : \lambda f.[S](\eta' f)}\,{<}}{S_\phi\backslash(S_\phi/NP_\phi) : [S](\rho'(\lambda p.p\ {*}completeness'))}\,{<}}{S_\phi : [S](\rho'(\lambda p.p\ {*}completeness'))([H](\theta'(\lambda x.{*}prove' x\ marcel')))}\,{<}$$

Invisible boundaries

- The majority of themes in utterances are null themes, unmarked by explicit boundary tones.
- The position of the theme-rheme boundary is usually ambiguous in these cases, as for example in (34).

(34) a. (I read a book about)$_{Theme}$ (COMPLETENESS)$_{Rheme}$
b. (I read)$_{Theme}$ (a book about COMPLETENESS)$_{Rheme}$
c. (I)$_{Theme}$ (read a book about COMPLETENESS)$_{Rheme}$
d. (I read a book about COMPLETENESS)$_{Rheme}$

- Steedman assumes that intermediate phrase L and H boundaries are indistinguishable from the null tone and may therefore be postulated anywhere there is no tone.

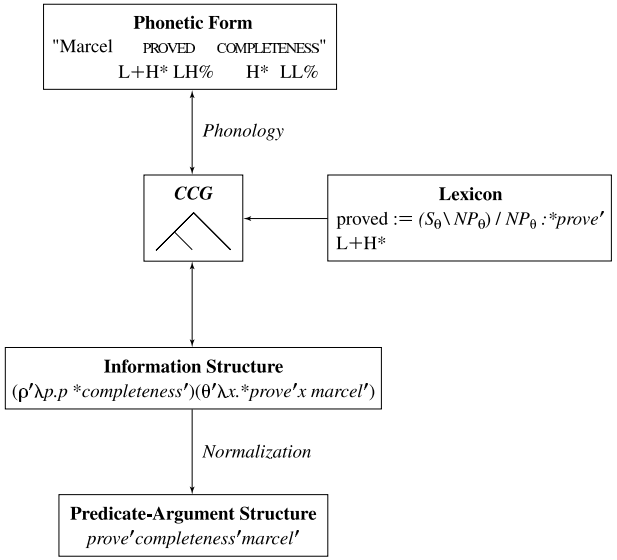
Invisible boundaries (cont.)

- Invisible boundaries can act as an edge of an unmarked theme.
- Undetectable boundaries are also allowed in other positions where there is no tone; for example, at the right-hand edge of an utterance-initial rheme followed by an unmarked theme.

(35) Q: Who proved completeness?
A: (MARCEL) (proved completeness).
H* L LL%

(36) I read a book about COMPLETENESS
L H* L L%
S/NP $S\$_\iota\backslash S\$_\eta$ $S_\rho\backslash(S_\rho/NP_\rho)$ $S\$_\iota\backslash S\$_\eta$ $(S\$_\phi\backslash S\$_\eta)\backslash(S\$_\iota\backslash S\$_\eta)$
S_ι/NP_ι $<$
$S\$_\phi\backslash S\$_\eta$ $<$
$S_\phi\backslash(S_\phi/NP_\phi)$ $<$
S_ϕ

Overview of Steedman's architecture



Open issues for Steedman's approach

- Steedman's approach requires continuous constituents since only adjacent material can be combined. This seems to incorrectly predict that information structure units must be continuous.
- Steedman's account seems to lack a restrictive theory of theme/rheme projection. How is projection of the rheme restricted, for example, from the subject onto the verbal projection, given that the subject and the verb can form a constituent in his approach? How can word order changes restrict projection?
- How can multiple focus (= rheme in Steedman's terminology) constructions be dealt with?
- Is there convincing motivation for the empty categories Steedman introduces for invisible boundary tones?

References

Baldridge, Jason and Geert-Jan Kruijff (2002). Coupling CCG with Hybrid Logic Dependency Semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, PA. http://www.aclweb.org/anthology/P02-1041.pdf.

Jackendoff, Ray (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Kruijff, Geert-Jan and Jason Baldridge (2004). Generalizing Dimensionality in Combinatory Categorial Grammar. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*. Geneva. http://www.cogsci.ed.ac.uk/~jmb/KruijffBaldridge-coling04.pdf.

Kruijff, Geert-Jan M. (2001). A Categorial-Modal Architecture of Informativity: Dependency Grammar Logic and Information Structure. Ph.D. thesis, Charles University, Prague, Czech Republic.

Steedman, Mark (2000a). Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry* 31(4), 649–689.

Steedman, Mark (2000b). *The Syntactic Process*. Cambridge, MA: MIT Press. Bradford Books.