1. Linguistics, Computation, and Modeling Human Language

Mark Steedman

Mar 13, 2017

School of informatics

Probabilistic Models for NLP with CCG

1. Linguistics, Computation, and Modeling Human Language
2. Combinatory Categorial Grammar for NLP
3. Wide Coverage Parsing with Combinatory Grammars
4. Robust Semantics for NLP

Prologue

- In the late 60's and the early 70's, linguists, psychologists and computational linguists saw themselves as engaged in the same project of understanding human language, using:
  - The formal theory of grammar proposed by Chomsky (1957, 1965);
  - The psycholinguistic theory of Miller *et al.* (1960); Miller (1967), as elaborated by Fodor *et al.* (1974);
  - The algorithmic theories of Thorne *et al.* (1968); Woods (1970).

Prologue

- Within a few years, this consensus fell apart:
  - Linguistic theory retreated behind the Competence-Performance distinction, claiming the cognitive inscrutability of the former;
  - Psycholinguists realized that linguistic theory made no strong predictions about processing difficulties, and either became agnostic about the relation of linguistic theory to mechanism, or went into connectionist denial;
  - Computational linguists realized that nothing that the other groups believed in was practically computable at the necessary scale and abandoned linguistic theory entirely in favor of Finite State Methods and Context Free Grammar.

- What went wrong?

Outline

- I: Chomsky (1957, 1965)
- II: Combinatory Categorial Grammar (CCG) as a Theory of Human Processing
- III: CCG as a Linguistic Theory
- IV: CCG and Incrementality in Human Sentence Processing
- V: Moral.

I: Chomsky's Definition of the Problem

- The Two Programs defined in *Syntactic Structures*:
  - Explanatory Adequacy: Identifying Complexity and Expressivity in the Theory of Grammar ;
  - Descriptive Adequacy: Capturing the phenomena of natural languages formally

- The Pessimistic Conclusions of *Aspects*:
  - To attain Explanatory Adequacy was impossibly difficult in the near term;
  - Descriptive Adequacy was susceptible to Cartesian (Euclidean) analysis using Transformations;.
  - Explanation would emerge from "significant generalizations" about observed constraints on transformations.

Some Misconceptions

- The linguists mistook the methodological priority of competence for a license to abdicate any responsibility for controling the degrees of freedom in the theory, compromising any claim to explanatory adequacy;[1]
- The psychologists assumed that the problem of performance was that there were at most two syntactic analyses of every sentence, and proceeded to construct surface-structure grammars of their own in terms of parsing preferences;
- The computational linguists became obsessed with the fact that there are actually hundreds, frequently thousands, and on occasion millions of syntactically well-formed analyses of sentences of even moderate length, focusing on the problem of search, at the expense of restricting grammar to finite-state or at most context-free power.

[1] Bizarrely, they identified explanatory adequacy in the theory with learnability in the limit.

Human and Computational NLP

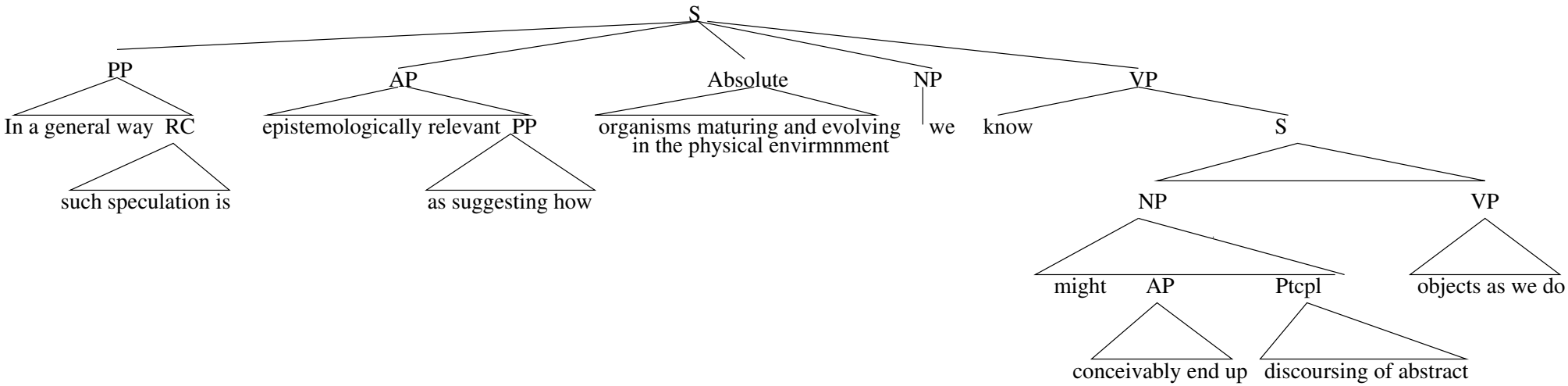
- No handwritten grammar ever has the coverage that is needed to read the daily newspaper. The grammars in our heads are huge

- Language is syntactically hugely ambiguous and it is hard to pick the best parse. Quite ordinary sentences of the kind you read every day routinely turn out to have hundreds and on occasion thousands of parses, albeit mostly semantically wildly implausible ones.

- High ambiguity and long sentences break exhaustive parsers.

What it's Really Like to be a Parser

- "In a general way such speculation is epistemologically relevant, as suggesting how organisms maturing and evolving in the physical environment we know might conceivably end up discoursing of abstract objects as we do." (Quine, 1960:123, via Abney, 1996):



Anatomy of a Natural Language Processor

- Every parser can be characterized by three elements:
  - A Grammar determined by the semantics (Regular, Context Free, Linear Indexed, etc.) and an associated automaton (Finite state, Push-Down, Extended Push-Down, etc.), together with the necessary working memories (stacks, registers, etc.);
  - A Search Algorithm (left-to-right etc., bottom-up etc.), etc.;
  - An Oracle, to resolve ambiguity and nondetermism (lexical, structural, etc.) on some criterion (statistical, semantic, etc.).
- The oracle can be used in two ways: either to actively limit the search space; or in the case of an all paths parser, to rank the results.
- In wide coverage parsing, we use it in the former way.

Competence and Performance

- Linguists (Chomsky 1957, *passim*), have always insisted on the methodological priority of "Competence" (the grammar that linguists study) and "Performance" (the mechanisms of language use).
- This makes sense: there are many possible parsers for each grammars.
- Nevertheless, Competence and Performance must have evolved as a single package, for what evolutionary edge does a parser without a grammar have, or a grammar without a parser?

(Although, since the evolution of language itself seems to have been essentially instantaneous, the package must have evolved for some other use, Steedman, 2002.)

Competence and Performance

- It follows that any theory that does not allow a one-to-one relation between grammatical and derivational constituency has some explaining to do.
- This observation suggests the following very strong assumption about the parser:
  - The Strict Competence Hypothesis: the parsing algorithm can only build structures that are licensed by the Competence Grammar as typable *constituents*.
- A corollary of SCH is that anything the parser shows evidence of building must be a constituent of competence grammar
- This includes the psychological oracle, which therfore pretty much has to be a generative model, derivable from the grammar.

Human Sentence Processing

- "Garden path" sentences are sentences which are grammatical, but which naive subjects fail to parse.
- Example (1a) is a garden path sentence, because the ambiguous word "sent" is analysed as a tensed verb:

(1) a. # The doctor sent for the patient died.
b. The flowers sent for the patient died.

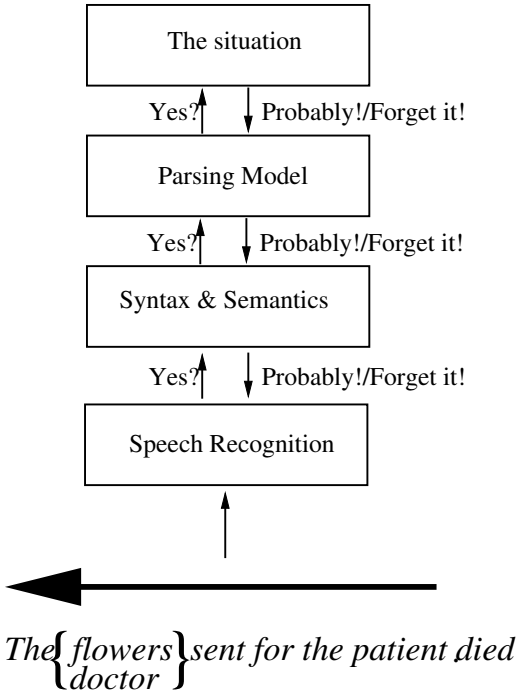
- However (1b) is not a garden path.
- So garden path effects are sensitive to something more than syntax (Bever 1970).

Human Sentence Processing

- "Something more" might be semantics/world-knowledge (or some proxy such as a probabilistic head-word dependency parsing model, or an RNN supertagger).
- They are even sensitive to referential context:
  - Crain and Steedman (1985) and Altmann and Steedman (1988) showed (simplifying somewhat) that if a context is established with two doctors, one of whom was sent for a patient, then the garden path effect is reversed.
- Whatever it is, the anomaly of "flowers" as a subject of "sent for" must have its effect before "the patient" is combined and the disambiguating main verb is encountered.
- If so, strict competence says that the main verb analysis of "The flowers sent for" must be a typable constituent.

The Architecture of the Processor

- This "weak" or "filtering" interaction requires incremental processing with a "cascade" architecture:



Requirements

- The requirements of incremental processsing, the strict competence condition, and syntax-semantics homomorphism are hard to satisfy simultaneously

II: CCG as a Theory of Human Performance

- CCG began as an attempt on explanatory adequacy via the idea that competence must be computationally grounded in performance.
- Such grounding immediately requires that the theory of grammar be polynomially decidable, to guarantee access to efficient divide-and-conquer algorithms such as CKY
- We started from the Harman/Gazdar insight that a great deal of the descriptive problem could be solved with context-free power, and Bill Woods's idea that the rest could mediated by a HOLD register that seemed to work like a stack.
- The idea was to do the work of both the PDA and the ATN HOLD register with the same stack (Ades and Steedman, 1982). (Cf. Joshi *et al.*, 1991; Kuhlmann *et al.*, 2015)

CCG as a Theory of Human Performance

- We also emphasized incremental syntactic and semantic processing
- We proposed a Bottom-Up Shift-Reduce architecture using a knowledge-rich parsing model to disambiguate categories and attachment.

The Paranoid Style in NLP

- We were immediately attacked by everybody:
  - For confusing performance with competence and not identifying the grammar in declaritive terms (the linguists);
  - For being not incremental enough (the psychologists);
  - For proliferating "spurious" syntactic ambiguity (the computational linguists).

- —and by everyone for believing in semantics

Our computers were also too small to do actually do any of this, and other than a few researchers in automatic speech processing (ASR) and machine translation (MT), none of us understood the role of statistical modeling.

III: CCG as Linguistic Theory

- CCG eschews language-specific syntactic rules like (4) for English.

(2) $S \rightarrow NP \ VP$
$VP \rightarrow TV \ NP$
$TV \rightarrow \{proved, found, met, \ldots\}$

- Instead, all language-specific syntactic information is *lexicalized*, via lexical entries like (5) for the English transitive verb:

(3) $\text{met} := (S\backslash NP)/NP$

- This syntactic "category" identifies the transitive verb as a function, and specifies the type and directionality of its arguments and the type of its result.

CCG as Linguistic Theory

- CCG eschews language-specific syntactic rules like (4) for English.

(4) $S \rightarrow NP \ VP$
$VP \rightarrow TV \ NP$
$TV \rightarrow \{proved, found, met, \ldots\}$

- Instead, all language-specific syntactic information is *lexicalized*, via lexical entries like (5) for the English transitive verb:

(5) $\text{met} := (S\backslash NP)/NP : \lambda x \lambda y.met\,x\,y$

- This syntactic "category" identifies the transitive verb as a function, and specifies the type and directionality of its arguments and the type of its result.

Type Raising as Case

- Type-raising in the form of case is a universal primitive of grammar

All noun-phrases (NP) like "Harry" are (polymorphically) type-raised.

- In Japanese and Latin this is the job of case morphemes like nominative *-ga* and *-us*.

- In English NPs are underspecified as to case, and must be disambiguated by the parsing model.

- Cf. the proposal of Vergnaud (1977/2006).

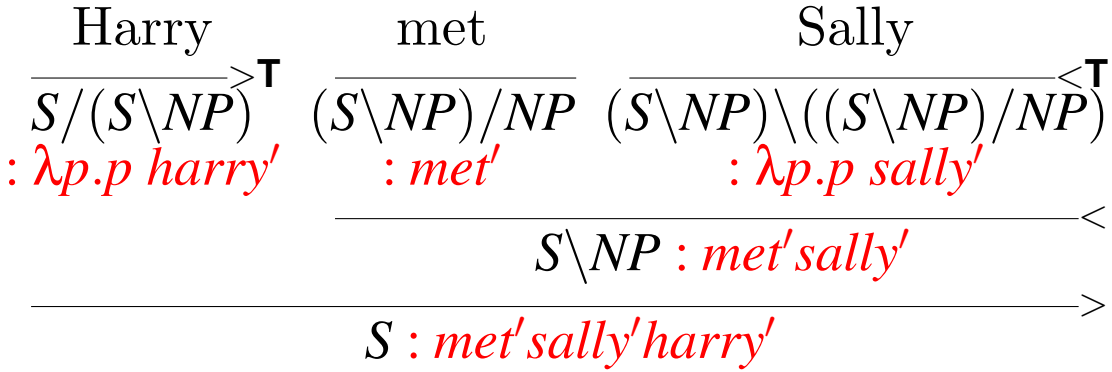
Syntactic Derivation

- (6)

$$\dfrac{\dfrac{\text{Harry}}{S/(S\backslash NP)}\,{>}\mathbf{T} \quad \dfrac{\dfrac{\text{met}}{(S\backslash NP)/NP} \quad \dfrac{\text{Sally}}{(S\backslash NP)\backslash((S\backslash NP)/NP)}\,{<}\mathbf{T}}{S\backslash NP}\,{<}}{S}\,{>}$$

"Surface Compositional" Semantics

- (7)



Relativization

- (8) that := $(N\backslash N)/(S/NP)$

(9) (The woman)

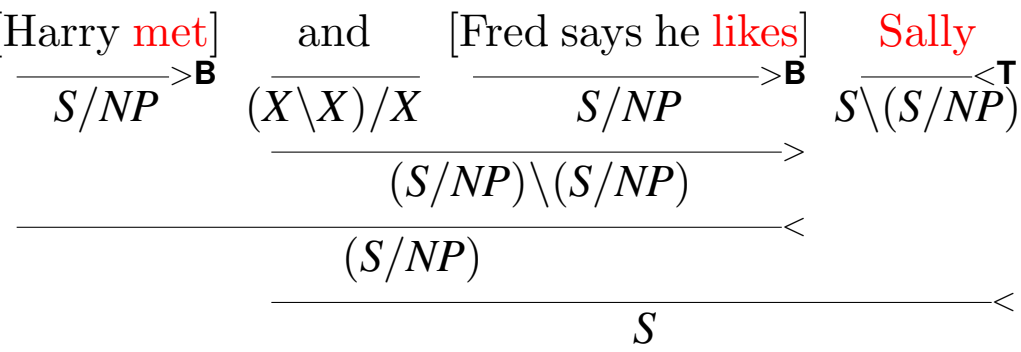
$$\dfrac{\dfrac{\text{that}}{(N\backslash N)/(S/NP)}\quad \dfrac{\dfrac{\text{Harry}}{S/(S\backslash NP)}\,{>}\mathbf{T}\quad \dfrac{\text{met}}{(S\backslash NP)/NP}}{S/NP}\,{>}\mathbf{B}}{N\backslash N}\,{>}$$

(10) (The woman)

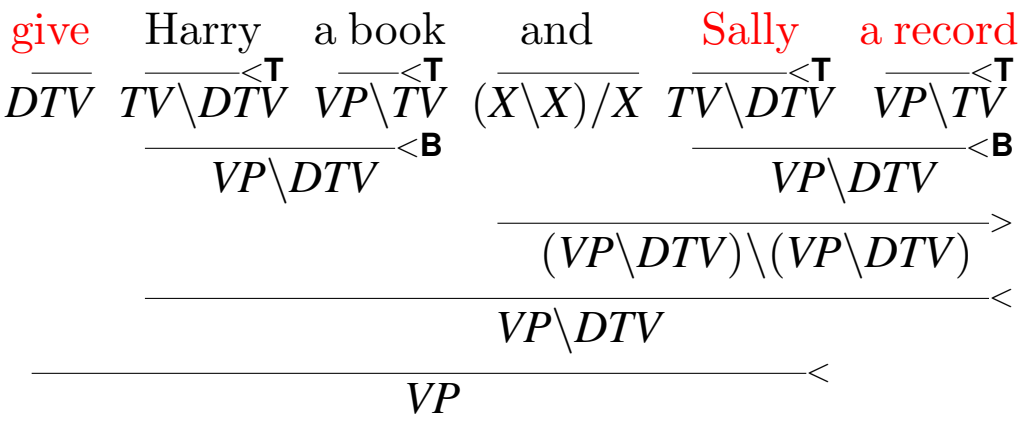
$$\dfrac{\dfrac{\text{that}}{(N\backslash N)/(S/NP)}\quad \dfrac{\dfrac{\dfrac{\text{Harry}}{S/(S\backslash NP)}\,{>}\mathbf{T}\quad \dfrac{\text{says}}{(S\backslash NP)/S}}{S/S}\,{>}\mathbf{B}\quad \dfrac{\dfrac{\text{he}}{S/(S\backslash NP)}\,{>}\mathbf{T}\quad \dfrac{\text{met}}{(S\backslash NP)/NP}}{S/NP}\,{>}\mathbf{B}}{S/NP}\,{>}\mathbf{B}}{N\backslash N}\,{>}$$

Coordination

- (11)



- (12)



Ross's Generalization

- The argument cluster coordination construction (12) is an example of a universal tendency for "deletion under coordination" to respect basic word order: in all languages, if arguments are on the left of the verb then argument clusters coordinate on the left, if arguments are to the right of the verb then argument clusters coordinate to the right of the verb (Ross 1970):

(13)

SVO:	*SO and SVO	SVO and SO
VSO:	*SO and VSO	VSO and SO
SOV:	SO and SOV	*SOV and SO

CCG reduces the linguists' MOVE and COPY/DELETE to adjacent MERGE

The Spurious Problem of "Spurious Ambiguity"

- (14)

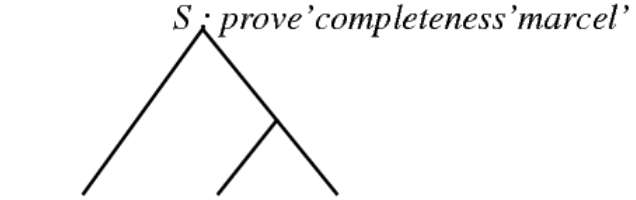
$$\dfrac{\dfrac{\dfrac{\text{Harry}}{S/(S\backslash NP) : \lambda p.p\ harry'}\,{>}\mathbf{T} \quad \dfrac{\text{met}}{(S\backslash NP)/NP : met'} \quad \dfrac{\text{Sally}}{S\backslash(S/NP) : \lambda p.p\ sally'}\,{<}\mathbf{T}}{S/NP : \lambda x.met'\ x\ harry'}\,{>}\mathbf{B}}{S : met'\ sally'\ harry'}\,{<}$$

The Spurious Problem of"Spurious" Ambiguity

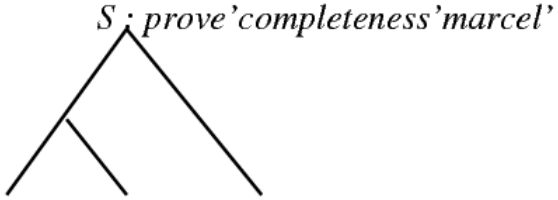
- Examples like the coordinate structures and relative clauses embody the claim that fragments like "Harry met", and "Sally a record", are constituents with the same standing as "met Sally".

- If such fragments can be constituent in right node raising, then they can be constituents of canonical sentences.

- Even such simple sentences are derivationally ambiguous:



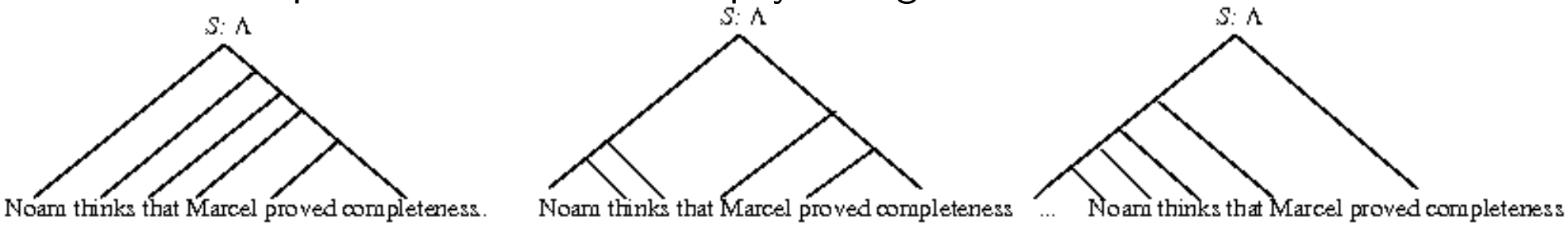
a. Marcel proved completeness



b. Marcel proved completeness

On So-called "Spurious" Ambiguity

- More complex sentences are multiplly ambiguous:



- This has been referred to (misleadingly) as "Spurious" ambiguity, since all the derivations have the same interpretation Λ.
- Interestingly, so called "spurious" constituents include most left prefixes.
- This means that a purely generative parsing model can potentially be incremental

Parsing in the Face of "Spurious Ambiguity"

- **All** grammars exhibit derivational ambiguity—even CFG.
- **Any** grammar that captures coordination at all will have the **same** derivational ambiguity as CCG.
- Use standard table-driven parsing methods such as CKY, with packed charts, where an entry is ruled admissible either by:
  - checking non-identity of **underlying** representation as table entries (Steedman 2000), rather than identity of derivation, or:
  - parsing normal-form derivations (Eisner 1996; Hockenmaier and Bisk 2010)

IV: CCG and Incrementality

- Most (but not all) left prefix substrings of sentences are typable constituents in CCG, for which alternative analyses can be compared using the parsing model
- The fact that (15a,b) involve the nonstandard constituent [The doctor sent for]$_{S/NP}$, means that constituent is also available for (15c,d)

(15) a. The patient that [the doctor sent for]$_{S/NP}$ died.

b. [The doctor sent for]$_{S/NP}$ and [The nurse attended]$_{S/NP}$ the patient who had complained of a pain.

c. #[The doctor sent for] $\left\{ \begin{array}{l} S/NP \\ (S/(S\backslash NP))/N \quad N \quad (N\backslash N)/NP \end{array} \right\}$ [the patient]$_{NP}$ died$_{S\backslash NP}$.

d. [The flowers sent for] $\left\{ \begin{array}{l} \#S/NP \\ (S/(S\backslash NP))/N \quad N \quad (N\backslash N)/NP \end{array} \right\}$ [the patient]$_{NP}$ died$_{S\backslash NP}$.

CCG and Incrementality

- (16) a. #[The doctor sent for the patient] $_S$ died$_{S\backslash NP}$.
  b. [The flowers sent for the patient died$_S$.

- Since the spurious constitutent [#The flowers sent for]$_{S/NP}$ is available in the chart, so that its low probability in comparison with the probabilities of the unreduced components can be detected (according to some "figure of merit" (Charniak *et al.* 1998) discounting the future), the garden path in (1b) is avoided,

Incrementality in Verb-final Languages

- If SO clusters in SOV languages can coordinate, as Ross observed, they must be constituents.
- If they are constituents, they can be constituents of canonical SOV sentences in languages like German and Japanese.
- If so, they too can support incremental parsing models for those languages under the Strict Competence Hypothesis
- There is abundant experimental evidence that sentence processing in verb-final languages is just as incremental as in English (Kamide and Mitchell, 1999; Kamide *et al.*, 2003a,b; Kazanina, 2016, *passim)*.

 —not to mention strong native speaker intuitions concerning incrementality in interpretation.

Wide-coverage Incremental CCG Parsing

- Existence of garden paths suggests human parsing is greedy and incremental
- The problem with greedy parsing is that the grammar is genuinely non-deterministic, prompting the use of lookahead and/or backtracking.
- Zhang and Clark (2011); Xu *et al.* (2014); Ambati *et al.* (2015) report partially incremental parsing algorithms for CCG that avoid backtracking using global linear parsing models, but use a lookahead of three words.
- Ambati (2016) reports a fully incremental version of his parser that eschews lookahead by using a narrow (16) beam, and constitutes a possible psycholinguistic model.

Moral

- It seems possible that we might be able to put NLP back together again.
- If so, there is more work to be done:
  - CCG must engage with the Minimalist linguists' aims of showing that the degrees of freedom in the theory are necessary and sufficient to capture the degrees of freedom in the syntactic data.
  - Psycholinguistics needs to engage with computational methods at the level of algorithms, rather than general principles like "top down" and "bottom up".
  - Computational Linguistics is soon going to have to lift its head above the level of Deep Learning applied to all the Low-hanging Fruit they've already shown can by captured by machine-learning.

References

Abney, Steven, 1996. "Statistical Methods and Linguistics." In Judith Klavans and Philip Resnik (eds.), *The Balancing Act*, Cambridge, MA: MIT Press. 1–26.

Ades, Anthony and Steedman, Mark, 1982. "On the Order of Words." *Linguistics and Philosophy* 4:517–558.

Altmann, Gerry and Steedman, Mark, 1988. "Interaction with Context During Human Sentence Processing." *Cognition* 30:191–238.

Ambati, Bharat Ram, 2016. *Transition-based Combinatory Categorial Grammar parsing for English and Hindi*. Ph.D. thesis, University of Edinburgh.

Ambati, Bharat Ram, Deoskar, Tejaswini, Johnson, Mark, and Steedman, Mark, 2015. "An Incremental Algorithm for Transition-based CCG Parsing." In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Denver, 53–63.

Bever, Thomas, 1970. "The Cognitive Basis for Linguistic Structures." In John Hayes (ed.), *Cognition and the Development of Language*, New York: Wiley. 279–362.

Charniak, Eugene, Goldwater, Sharon, and Johnson, Mark, 1998. "Edge-Based Best-First Chart Parsing." In *Proceedings of the 6th Workshop on Very Large Corpora, Montreal, August*. 127–133.

Chomsky, Noam, 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam, 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Crain, Stephen and Steedman, Mark, 1985. "On Not Being Led up the Garden Path: The Use of Context by the Psychological Parser." In David Dowty, Lauri Kartunnen, and Arnold Zwicky (eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, Cambridge: Cambridge University Press. 320–358.

Eisner, Jason, 1996. "Efficient Normal-Form Parsing for Combinatory Categorial Grammar." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA*. San Francisco: Morgan Kaufmann, 79–86.

Fodor, Jerry A., Bever, Thomas, and Garrett, Merrill, 1974. *The Psychology of Language*. New York: McGraw-Hill.

Hockenmaier, Julia and Bisk, Yonatan, 2010. "Normal-Form Parsing for Combinatory Categorial Grammars with Generalized Composition and Type-Raising." In *Proceedings of the 23nd International Conference on Computational Linguistics*. Beijing, 465–473.

Joshi, Aravind, Vijay-Shanker, K., and Weir, David, 1991. "The Convergence of Mildly Context-Sensitive Formalisms." In Peter Sells, Stuart Shieber, and Tom Wasow (eds.), *Processing of Linguistic Structure*, Cambridge, MA: MIT Press. 31–81.

Kamide, Yuki, Altmann, Gerry, and Haywood, Sarah, 2003a. "The Time-Course of Prediction in Incremental Sentence Processing: Evidence from Anticipatory Eye Movements." *Journal of Memory and Language* 49:133–156.

Kamide, Yuki and Mitchell, Don, 1999. "Incremental Pre-head Attachment in Japanese Parsing." *Language and Cognitive Processes* 11:631–662.

Kamide, Yuki, Scheepers, Christoph, and Altmann, Gerry TM, 2003b. "Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English." *Journal of psycholinguistic research* 32:37–55.

Kazanina, Nina, 2016. "Predicting Complex Syntactic Structure in Real Time: Processing of Negative Sentences in Russian." *The Quarterly Journal of Experimental Psychology* :1–19.

Kuhlmann, Marco, Koller, Alexander, and Satta, Giorgio, 2015. "Lexicalization and Generative Power in CCG." *Computational Linguistics* 41:187–219.

Miller, George, 1967. "Project Grammarama." In George Miller (ed.), *The Psychology of Communication*, New York: Basic Books. 125–187.

Miller, George, Galanter, Eugene, and Pribram, Karl, 1960. *Plans and the Structure of Behavior*. New York: Holt.

Quine, Willard Van Orman, 1960. *Word and Object*. Cambridge, MA: MIT Press.

Ross, John Robert, 1970. "Gapping and the Order of Constituents." In Manfred Bierwisch and Karl Heidolph (eds.), *Progress in Linguistics*, The Hague: Mouton. 249–259.

Steedman, Mark, 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.

Steedman, Mark, 2002. "Plans, Affordances, and Combinatory Grammar." *Linguistics and Philosophy* 25:723–753.

Thorne, James, Bratley, Paul, and Dewar, Hamish, 1968. "The Syntactic Analysis of English by Machine." In Donald Michie (ed.), *Machine Intelligence*, Edinburgh: Edinburgh University Press, volume 3.

Vergnaud, Jean Roger, 1977/2006. "Letter to Noam Chomsky and Howard Lasnik on "Filters and Control," April 17, 1977." In Robert Freidin and Howard Lasnik (eds.), *Syntax: Critical Concepts in Linguistics*, Oxford: Blackwell. 21–34.

Woods, William, 1970. "Transition Network Grammars for Natural Language Analysis." *Communications of the Association for Computing Machinery* 18:264–274.

Xu, Wenduan, Clark, Stephen, and Zhang, Yue, 2014. "Shift-Reduce CCG Parsing with a Dependency Model." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, MD, 218–227.

Zhang, Yue and Clark, Stephen, 2011. "Shift-Reduce CCG Parsing." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR: ACL, 683–692.