Introduction to Cognitive Science: Notes

IX: Human and Computational NLP

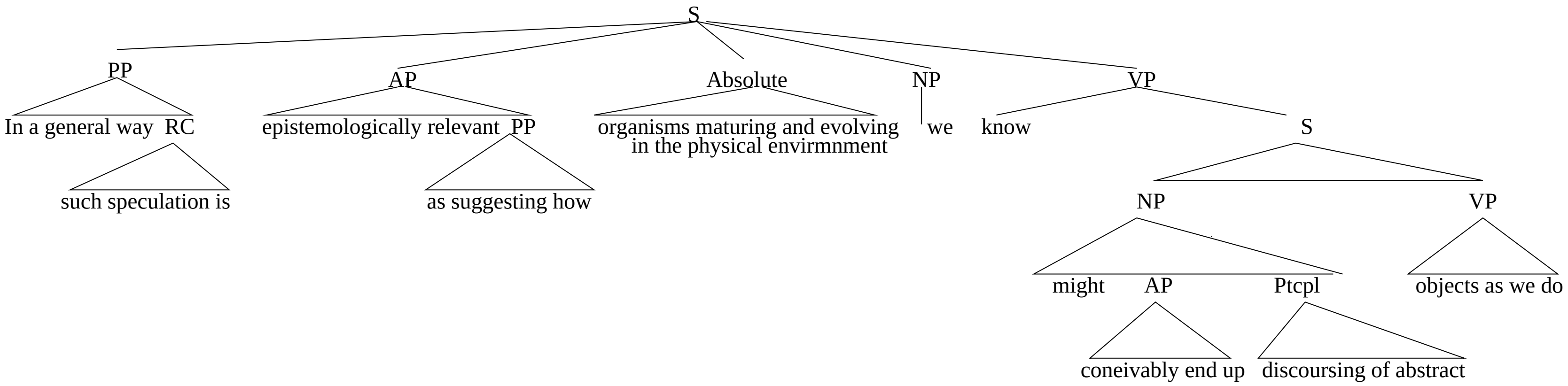
- Readings for this section: Pereira 2000; Altmann 1998.

IX: Human and Computational NLP

- No handwritten grammar ever has the coverage that is needed to read the daily newspaper.
- Language is syntactically highly 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.

For Example:

- “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).
- —yields the following (from Abney 1996), among many other horrors:



The Anatomy of a Parser

- Every parser can be identified by three elements:
  - A Grammar (Regular, Context Free, Linear Indexed, etc.) and an associated automaton (Finite state, Push-Down, Embedded Push-Down, etc.);
  - A search Algorithm characterized as left-to-right (etc.), bottom-up (etc.), and the associated working memories (etc.);
  - An Oracle, to resolve ambiguity.
- 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 have to use it in the former way.

Competence and Performance

- Chomsky (1957, *passim*), has always insisted on the methodological independence of “Competence” (the grammar that linguists study) and “Performance” (the mechanisms of language use).
- This makes sense: there are many more performance mechanisms than there are grammatical levels, and for any sentence there are many ways of uttering it.
- Nevertheless, Competence and Performance must have evolved as a single package, for what is a parser without a grammar, or a grammar without a parser/generator?

Human Sentence Processing

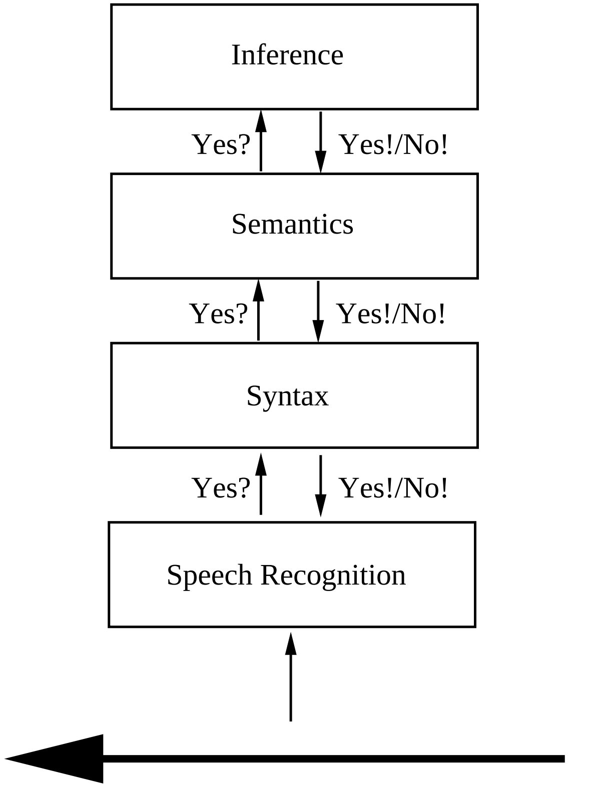
- “Garden path” sentences are sentences which are grammmatrical, but for which naive subjects fail to parse.
- Example (1a) is a garden path sentence, because the ambiguous word “sent” is analysed as a tensed verb:

(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 semantic content and pragmatic knowledge, (Bever 1970) and even to context (Altmann and Steedman 1988).

The Architecture of the Human Sentence Processor

- This requires a “cascade” architecture:



$$\textit{The}\left\{\begin{array}{l}\textit{flowers}\\ \textit{doctor}\end{array}\right\}\textit{sent for the patient died.}$$

Human Sentence Processing

- This architecture embodies Incremental Fine-grain Parallel, “Weakly Interactive” Parsing
- The “Weak” interaction with semantics is where syntax proposes interpretations, and semantics and pragmatics and inference in context then rank them for plausibility (Crain, Altmann, et al.).
- **Contexts:**

  A burglar broke into a bank carrying some dynamite.
  He planned to blow open a safe . . .

  - **NP-attachment-supporting continuation**:

    . . . Once inside he saw that there was a safe with a new lock and a safe with an old lock.
  - **VP-attachment-supporting continuation**:

    . . . Once inside he saw that there was a safe with a new lock and a strongbox with an old lock.

Human Sentence Processing

- **Target Sentences:**
  - **NP-attached target**:
    The burglar blew open the safe with the new lock and made off with the loot.
  - **VP-attached target**:
    The burglar blew open the safe with the dynamite and made off with the loot.

Weakly Interactive Parsing (contd.)

- If garden paths are under the control of context, why do the classic garden path sentences garden path in the *neutral* or “null” context?
  - Because the null context *isn’t* neutral. It is instead the *simplest* context compatible with the sentence being processed.
  - It is simpler to accomodate one safe than more than one, one horse rather than several horses, etc.
  - In the case of examples like *The horse raced past the barn fell* there are even more presuppositions to accomodate—like their being an activity which made one of the horses race alonfg a particular path.
- Can we afford to implement weak semantically interactive parsing in practice?
  - Yes, but only if we can model the knowledge of the domain completely.
  - Therefore, not for tasks like parsing the daily newspaper or arbitrarily-chosen web-pages.

Weak Interaction and Competence Grammar

- It is interesting that CCG’s unorthodox approach to syntactic constituency means that most left prefix substrings of sentences are typable constituents, complete with an interpretation.

- For example, the fact that (2a,b) involve the nonstandard constituent [The doctor sent for]$_{S/NP}$, means that constituent is also available for the canonical sentence (2c)

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

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

c. [The doctor sent for]$_{S/NP}$ the patient.

The Strict Competence Hypothesis

- This means that the spurious constitutent [#The flowers sent for]$_{S/NP}$ is also available with an interpretation, so that its semantic anomaly can be detected via the weak or filtering interaction, and the garden path in (1b) avoided, even under the following very strong assumption about the parser:
- The Strict Competence Hypothesis: the parser only builds structures that are licensed by the Competence Grammar as typable *constituents*.
- This is an attractive hypothesis, because it allows Competence Grammar and Performance Parser/Generator to evolve as a package deal, with parsing completely transparent to grammar.
- But is such a simple parser possible? We need to look at some real-life parsing programs.

Wide Coverage Parsing: the State of the Art

- Early attempts to model parse probability by attaching probabilities to rules of CFG performed poorly.
- Great progress as measured by the ParsEval measure has been made by combining statistical models of headword dependencies with CF grammar-based parsing (Collins 1997; Charniak 2000; Bod 2001)
- However, the ParsEval measure is very forgiving. Such parsers have until now been based on highly overgenerating context-free covering grammars. Analyses depart in important respects from interpretable structures.
- In particular, they fail to represent the long-range "deep" semantic dependencies that are involved in relative and coordinate constructions, as in $A\ company_i\ that_i\ I\ think\ IBM\ bought_i$, and $IBM_i\ bought_{i,j}\ and\ sold_{i,j}\ Lotus_j$.

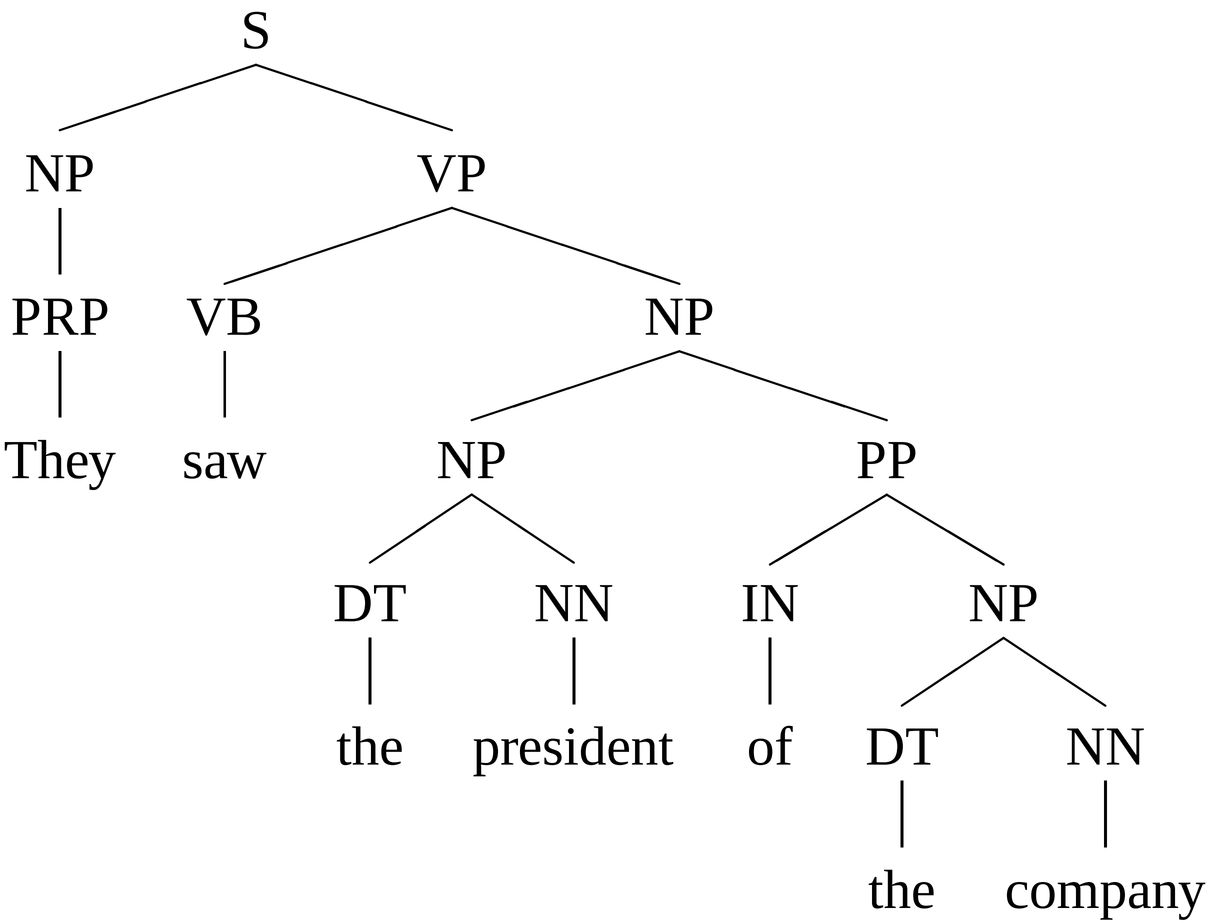
Statistical Models for Wide-Coverage Parsers

- There are two kinds of statistical models:
  - Generative models directly represent the probabilities of the rules of the grammar, such as the probability of the word *eat* being transitive, or of it taking a nounphrase headed by the word *integer* as object.
  - Discriminative models compute probability for whole parses as a function of the product of a number of weighted features, like a Perceptron. These features typically include those of generative models, but can be anything.

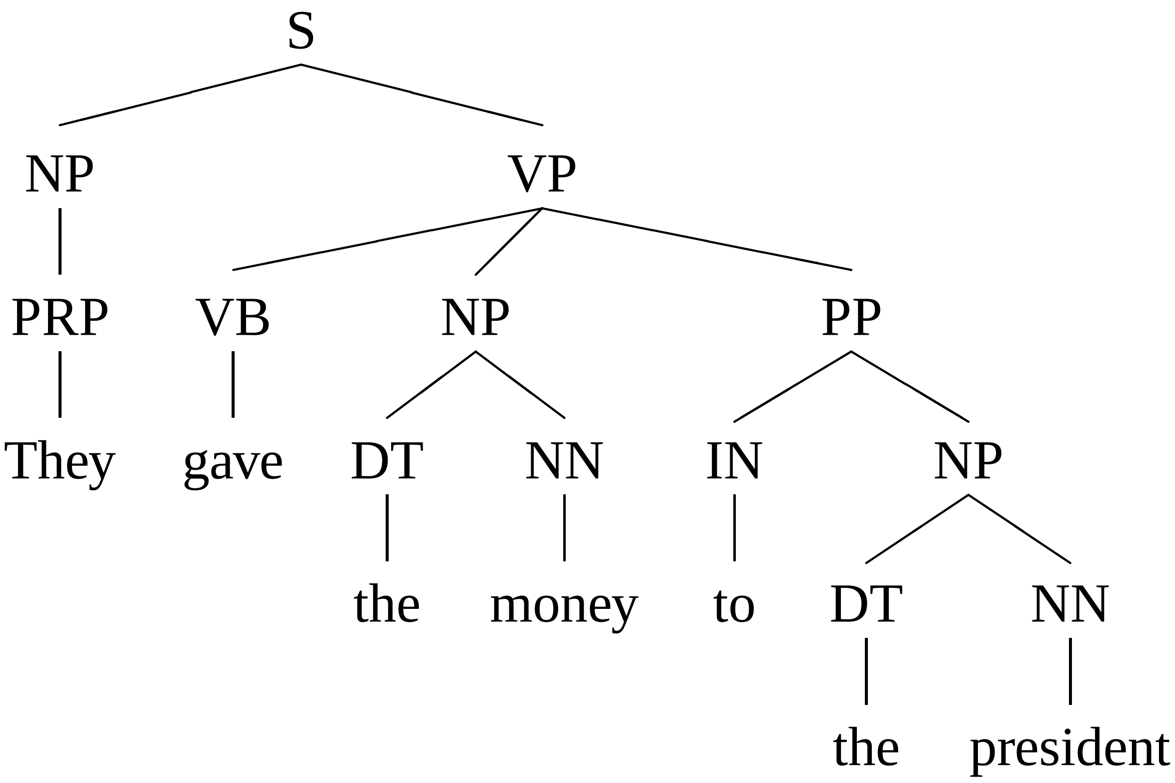
Selecting the best parse

- Basic PCFGs (for parse selection) have a number of problems:
  - They ignore lexical information (such as the likelihood of the verb *find* occuring as an intransitive verbphrase).
  - Are biased towards short flat parses.
  - Make unwarranted independence assumptions.
  - Counts are usually low, so it is important to smooth probabilities.

Selecting the best parse

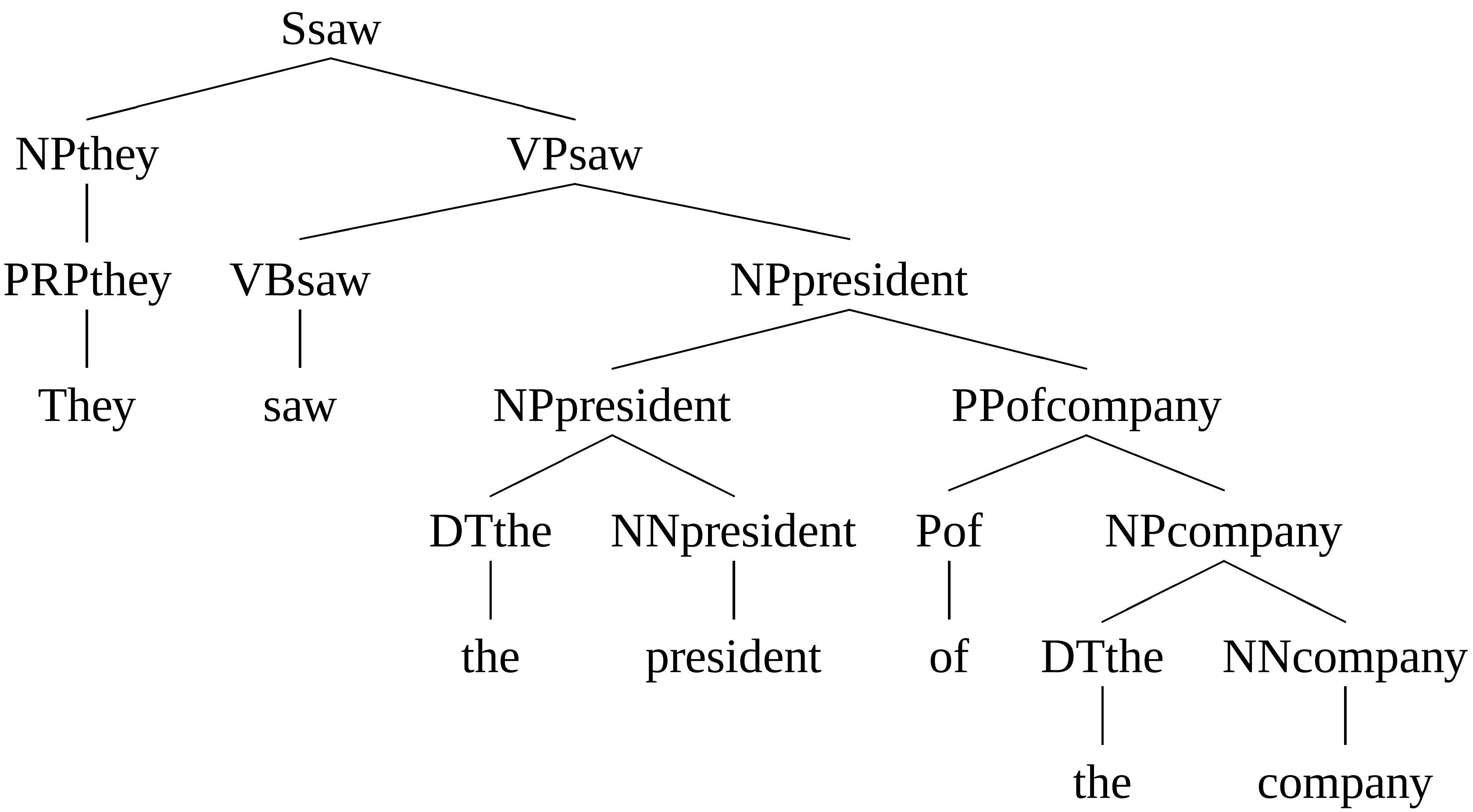


Selecting the best parse



Selecting the best parse

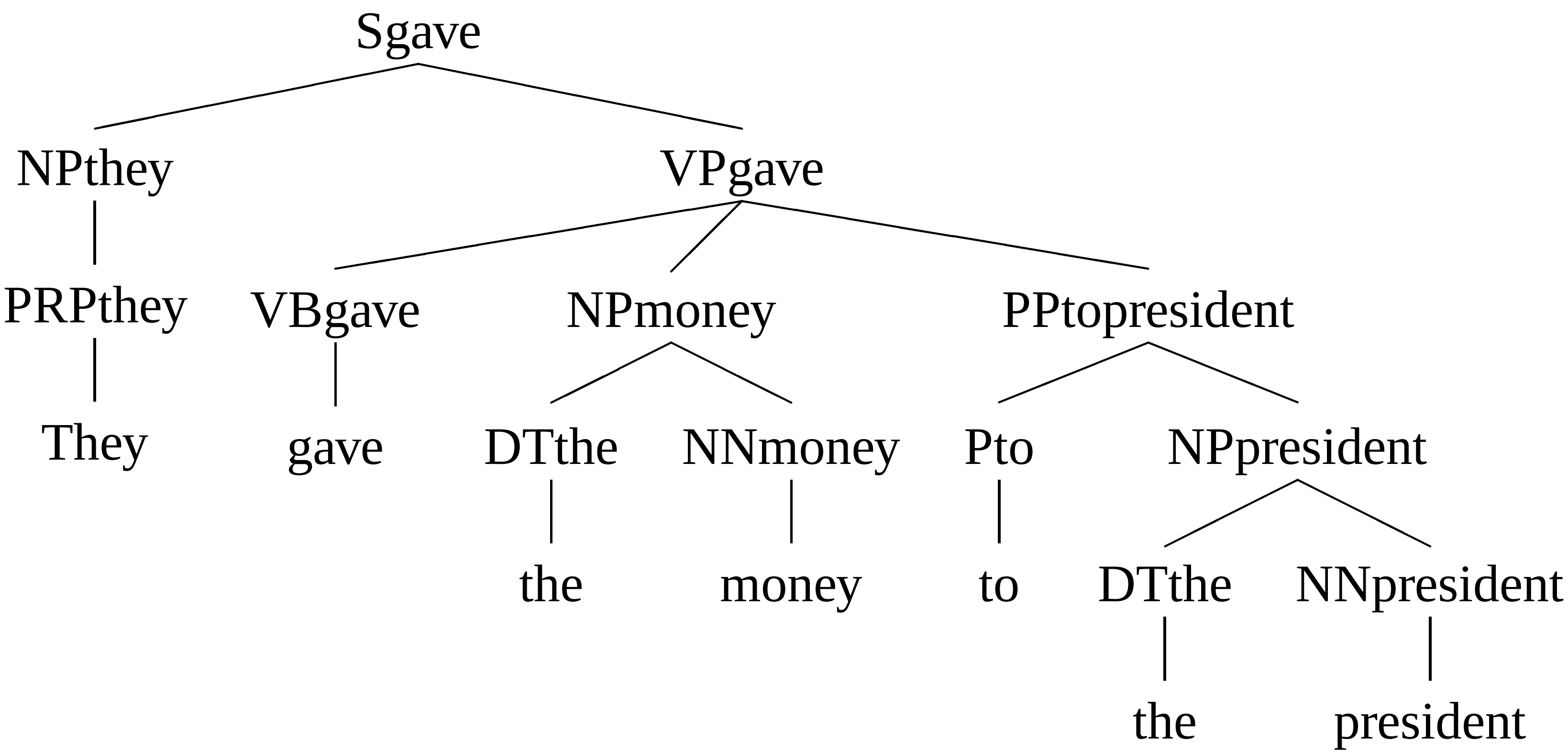
Lexicalisation:



(Magerman 1995, Collins 1997, Charniak 1997):

Selecting the best parse

Lexicalisation:



Why Lexicalize the Model?

- To assign probabilities to such trees, we need to be more careful of our independence assumptions

- A treebank:

```
[S(grows),[NP, grass],[VP,grows]]
[S(grows),[NP, grass],[VP,grows]]
[S(grows),[NP, rice],[VP,grows]]
[S(grow),[NP, bananas],[VP,grow]]
```

Why Lexicalize the Model? (Contd.)

- A naive PCFG is unsound!

Rule A → α					Count	P ≈ Rel.Frequency|A
r1	S	→	NP	VP	4	1
r2	NP	→	rice		1	1/4
r3	NP	→	grass		2	1/2
r4	NP	→	bananas		1	1/4
r5	VP	→	grows		3	3/4
r6	VP	→	grow		1	1/4

- P(S[NP[grass]VP[grows]]) = 1/2*3/4*1 = 3/8
- P(S[NP[rice]VP[grows]]) = 1/4*3/4*1 = 3/16
- P(S[NP[bananas]VP[grow]]) = 1/4*1/4*1 = 1/16 total Z = 5/8

Why Lexicalize the Model? (Contd.)

- Normalization with respect to Z doesn’t help:
- P(S[NP[grass]VP[grows]]) = 1/2*3/4*1*8/5 = 3/5
- P(S[NP[rice]VP[grows]]) = 1/4*3/4*1*8/5 = 3/10
- P(S[NP[bananas]VP[grow]]) = 1/4*1/4*1*8/5 = 1/10 total = 1
- These probabilities are still wrong by inspection.
- The problem is the independence assumption: there are non-local dependencies.

Why Lexicalize the Model? (Contd.)

- To build such trees, we separate the grammar into rules which say how heads are passed, and tables of dependency probabilities:
- A lexicalized PCFG:

Rule $A \longrightarrow \alpha$					Count	$P \approx$ Rel.Frequency$\|A$
r0	START	$\longrightarrow$	S(grows)		3	3/4
r0	START	$\longrightarrow$	S(grow)		1	1/4
r1	S(H2)	$\longrightarrow$	NP(H1)	VP(H2)	4	1
r2	NP(rice)	$\longrightarrow$	rice		1	1
r3	NP(grass)	$\longrightarrow$	grass		2	1
r4	NP(bananas)	$\longrightarrow$	bananas		1	1
r5	VP(grows)	$\longrightarrow$	grows		3	1
r6	VP(grow)	$\longrightarrow$	grow		1	1

Why Lexicalize the Model? (Contd.)

- $P([S,[NP,grass][VP,grows]])$ =
  $P(S(grows)|START) * P(r1|S(grows)) * P(VP(grows)|S(grows),r1,2) *$
  $P(NP(grass)|S(grows),r1,1)$
  $= 3/4 * 1 * 1 * 2/3 \qquad = 1/2$

- $P([S,[NP,rice][VP,grows]])$ =
  $P(S(grows)|START) * P(r1|S(grows)) * P(VP(grows)|S(grows),r1,2) *$
  $P(NP(rice)|S(grows),r1,1)$
  $= 3/4 * 1 * 1 * 1/3 \qquad = 1/4$

- $P([S,[NP,bananas][VP,grow]])$ =
  $P(S(grow)|START) * P(r1|S(grow)) * P(VP(grow)|S(grow),r1,2) *$
  $P(NP(bananas)|S(grow),r1,1)$
  $= 1/4 * 1 * 1 * 1 \qquad = 1/4$

- These probabilities are correct by observation. (In general they need to be normalized by length of derivation.)

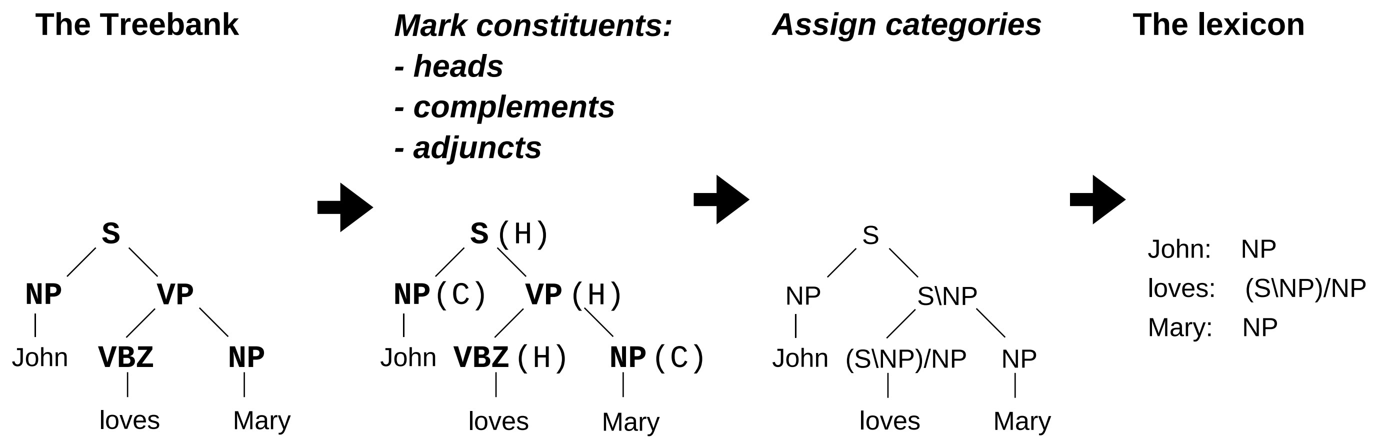
- Head dependencies are also getting us the effect of number agreement.

Head-dependencies as Oracle

- Head-dependency-Based Statistical Parser Optimization works because it approximates an oracle using semantics and real world inference.
- Its probably as close as we will get to the real thing for the foreseeable future.
- In fact, the knowledge- and inference- based psychological oracle may be much more like a probabilistic relational model than like traditional logicist representations, especially if embedded in associative knowledge representations, augmented by ontologies and integrated with a dynamic context model.
- Many context-free processing techniques generalize to the mildly context sensitive class.
- The “nearly context free” grammars such as LTAG and CCG—the least expressive generalization of CFG known—have been treated by Xia (1999), Hockenmaier and Steedman (2002), and Clark and Curran (2004).

Supervised CCG Induction by Machine

- Extract a CCG lexicon from the Penn Treebank: Hockenmaier and Steedman (2002), Hockenmaier (2003) (cf. Buszkowski and Penn 1990; Xia 1999).



- This trades lexical types (500 against 48) for rules (around 3000 instantiated binary combinatory rule types against around 12000 PS rule types) with standard Treebank grammars.

Supervised CCG Induction: Full Algorithm

-
  ```
  foreach tree T:
  preprocessTree(T);
  preprocessArgumentCluster(T);
  determineConstituentType(T);
  makeBinary(T);
  percolateTraces(T);
  assignCategories(T);
  treatArgumentClusters(T);
  cutTracesAndUnaryRules(T);
  ```

- The resulting treebank is somewhat cleaner and more consistent, and is offered for use in inducing grammars in other expressive formalisms. It was released in June 2005 by the Linguistic Data Consortium with documentation and can be searched using t-grep.

Overall Dependency Recovery

	LP	LR	UP	UR	cat
Clark et al. 2002	81.9	81.8	90.1	89.9	90.3
Hockenmaier 2003	84.3	84.6	91.8	92.2	92.2
Log-linear	**86.6**	**86.3**	**92.5**	**92.1**	**93.6**
Hockenmaier (POS)	83.1	83.5	91.1	91.5	91.5
Log-linear (POS)	**84.8**	**84.5**	**91.4**	**91.0**	**92.5**

Table 1: Dependency evaluation on Section 00 of the Penn Treebank

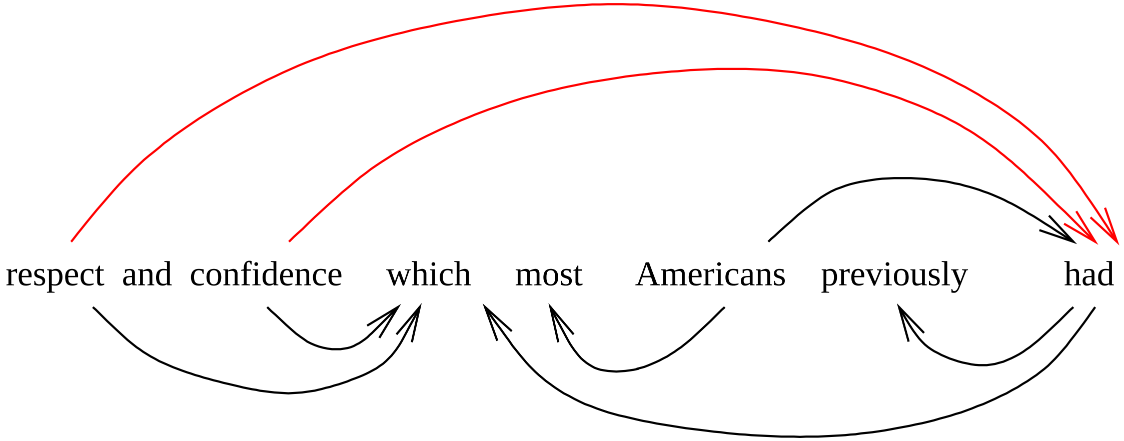
- To maintain comparability to Collins, Hockenmaier (2003) did not use a Supertagger, and was forced to use beam-search. With a Supertagger front-end, the Generative model might well do as well as the Log-Linear model. We have yet to try this experiment.

Log-Linear Overall Dependency Recovery

- The C&C parser has state-of-the-art dependency recovery.
- The C&C parser is very fast ($\approx$ 30 sentences per second)
- The speed comes from highly accurate supertagging which is used in a "Best-First increasing" mode (Clark and Curran 2004), and behaves as an "almost parser" (Bangalore and Joshi 1999).
- CCG almost-parsing is why Zettlemoyer and Collins do so well on a small not very ambiguous corpus without having a parser model at all (see next lecture).
- It has been ported to the TREC QA task (Clark *et al*. 2004), and applied to the entailment QA task (Bos *et al*. 2004), using automatically built logical forms.

Recovering Deep or Semantic Dependencies

Clark *et al.* (2002)



respect and confidence which most Americans previously had

lexical_item	category	slot	head_of_arg
which	$(NP_X\backslash NP_{X,1})/(S[dcl]_2/NP_X)$	2	*had*
which	$(NP_X\backslash NP_{X,1})/(S[dcl]_2/NP_X)$	1	*confidence*
which	$(NP_X\backslash NP_{X,1})/(S[dcl]_2/NP_X)$	1	*respect*
had	$(S[dcl]_{had}\backslash NP_1)/NP_2$	2	*confidence*
had	$(S[dcl]_{had}\backslash NP_1)/NP_2$	2	*respect*

Full Object Relatives in Section 00

- 431 sentences in WSJ 2-21, 20 sentences (24 object dependencies) in Section 00. 1. Commonwealth Edison now faces an additional court-ordered *refund* on its summerwinter rate differential collections *that* the Illinois Appellate Court has *estimated* at DOLLARS.

  2. Mrs. Hills said many of the 25 *countries that* she *placed* under varying degrees of scrutiny have made genuine progress on this touchy issue.

  √ 3. It's the petulant complaint of an impudent *American whom* Sony *hosted* for a year while he was on a Luce Fellowship in Tokyo – to the regret of both parties.

  √ 4. It said the *man*, *whom* it did not *name*, had been found to have the disease after hospital tests.

  5. Democratic Lt. Gov. Douglas Wilder opened his gubernatorial battle with Republican Marshall Coleman with an abortion *commercial* produced by Frank Greer *that* analysts of every political persuasion *agree* was a tour de force.

  6. Against a shot of Monticello superimposed on an American flag, an announcer talks about the strong *tradition* of freedom and individual liberty *that* Virginians have *nurtured* for generations.

  √ 7. Interviews with analysts and business people in the U.S. suggest that Japanese capital may produce the economic *cooperation that* Southeast Asian politicians have *pursued* in fits and starts for decades.

  8. Another was Nancy Yeargin, who came to Greenville in 1985, full of the *energy* and *ambitions that* reformers wanted to *reward*.

  9. Mostly, she says, she wanted to prevent the *damage* to self-esteem *that* her low-ability students would *suffer* from doing badly on the test.

  √ 10. Mrs. Ward says that when the cheating was discovered, she wanted to avoid the morale-damaging public *disclosure that* a trial would *bring*.

  √ 11. In CAT sections where students' knowledge of two-letter consonant sounds is tested, the authors noted that

Scoring High concentrated on the same *sounds that* the test *does* – to the exclusion of other *sounds that* fifth graders should *know*.

√ 12. Interpublic Group said its television programming *operations* – *which* it *expanded* earlier this year – agreed to supply more than 4,000 hours of original programming across Europe in 1990.

13. Interpublic is providing the programming in return for advertising *time*, *which* it *said* will be valued at more than DOLLARS in 1990 and DOLLARS in 1991.

√ 14. Mr. Sherwood speculated that the *leeway that* Sea Containers *has* means that Temple would have to substantially increase their bid if they're going to top us.

√ 15. The Japanese companies bankroll many small U.S. companies with promising products or ideas, frequently putting their money behind *projects that* commercial banks won't *touch*.

√ 16. In investing on the basis of future transactions, a role often performed by merchant banks, trading companies can cut through the *logjam that* small-company owners often *face* with their local commercial banks.

17. A high-balance *customer that* banks *pine for*, she didn't give much thought to the rates she was receiving, nor to the fees she was paying.

√ 18. The events of April through June damaged the *respect* and *confidence which* most Americans previously *had* for the leaders of China.

√ 19. He described the situation as an escrow *problem*, a timing *issue*, *which* he *said* was rapidly rectified, with no losses to customers.

√ 20. But Rep. Marge Roukema (R., N.J.) instead praised the House's acceptance of a new youth training wage, a *subminimum that* GOP administrations have *sought* for many years.

Cases of object extraction from a relative clause in 00; the extracted object, relative pronoun and verb are in italics; sentences marked with a √ are cases where the parser correctly recovers all object dependencies

References

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

Altmann, Gerry, 1998. “Ambiguity in Sentence Processing.” *Trends in Cognitive Sciences* 2:146–152.

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

Bangalore, Srinivas and Joshi, Aravind, 1999. “Supertagging: An Approach to Almost Parsing.” *Computational Linguistics* 25:237–265.

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

Bod, Rens, 2001. “What is the Minimal Set of Fragments that Achieves Maximal Parse Accuracy?” In *Proceedings of the 39th Meeting of the ACL*. Toulouse, France.

Bos, Johan, Clark, Stephen, Steedman, Mark, Curran, James R., and Hockenmaier, Julia, 2004. “Wide-Coverage Semantic Representations from a CCG Parser.” In *Proceedings of the 20th International Conference on Computational Linguistics (COLING ’04), Geneva*. ACL, 1240–1246.

Buszkowski, Wojciech and Penn, Gerald, 1990. “Categorial Grammars Determined from Linguistic Data by Unification.” *Studia Logica* 49:431–454.

Charniak, Eugene, 2000. “A Maximum-Entropy-Inspired Parser.” In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, 132–139.

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

Clark, Stephen and Curran, James R., 2004. “Parsing the WSJ using CCG and Log-Linear Models.” In *Proceedings of the 42nd Meeting of the ACL*. Barcelona, Spain, 104–111.

Clark, Stephen, Hockenmaier, Julia, and Steedman, Mark, 2002. “Building Deep

Dependency Structures with a Wide-Coverage CCG Parser.” In *Proceedings of the 40th Meeting of the ACL*. Philadelphia, PA, 327–334.

Clark, Stephen, Steedman, Mark, and Curran, James R., 2004. “Object-Extraction and Question-Parsing Using CCG.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, 111–118.

Collins, Michael, 1997. “Three Generative Lexicalized Models for Statistical Parsing.” In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid*. San Francisco, CA: Morgan Kaufmann, 16–23.

Hockenmaier, Julia, 2003. *Data and models for statistical parsing with CCG*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Hockenmaier, Julia and Steedman, Mark, 2002. “Acquiring Compact Lexicalized Grammars from a Cleaner Treebank.” In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 1974–1981.

Pereira, Fernando, 2000. “Formal Grammar and Information Theory: Together Again?” *Philosophical Transactions of the Royal Society* 385:1239–1253.

Quine, Willard van Ormond, 1960. *Word and Object*. Cambridge MA: MIT Press.

Xia, Fei, 1999. “Extracting Tree Adjoining Grammars from Bracketed Corpora.” In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*.