Sentence Processing
Lecture 5
Introduction to Psycholinguistics

Matthew W. Crocker
Pia Knoeferle

Department of Computational Linguistics
Saarland University

Reading

Altmann, G. Ambiguity in Sentence Processing. *Trends in Cognitive Sciences*, 2:4, 1998.

- How do the accounts Altmann discusses relate to the notion of linguistic modularity?
- What kinds of information is used during processing?
- We will return later in the course to:
  - theories of ambiguity resolution later
  - connectionist and constraint-based processing models

Next lecture: Experimental Methods II (PK)

Theories of Sentence Processing

Structure-based theories

- Disambiguation based on structural heuristics

Grammar-based theories

- Preferred structure based on grammatical principles

Experience-based theories

- Structural preferences are based on prior experience

Interactive accounts

- Disambiguation draws on diverse knowledge sources

Resources-based accounts

- Preferred structure involves the least resources

Linking Hypotheses

Relate the theory/model to some observed measure

- Typically impossible to predict measures completely

Theories of parsing typically determine …

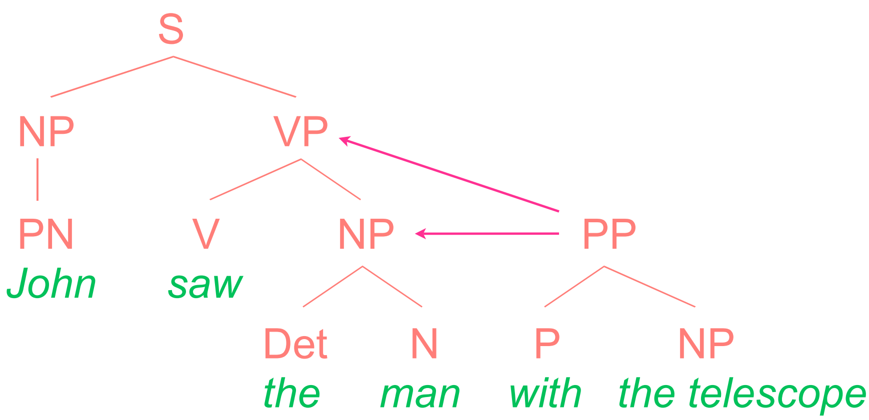
- what **mechanism** is used to construct interpretations?
- which **information** sources are used by the mechanism?
- which **representation** is preferred/constructed when ambiguity arises?

Linking Hypothesis:

- Preferred sentence structures should have faster reading times in the disambiguating region than dispreferred

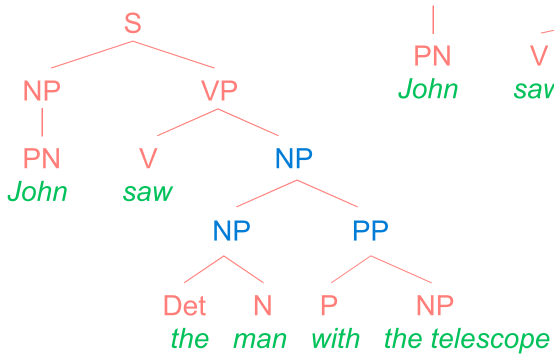
The Garden Path Theory (Frazier)

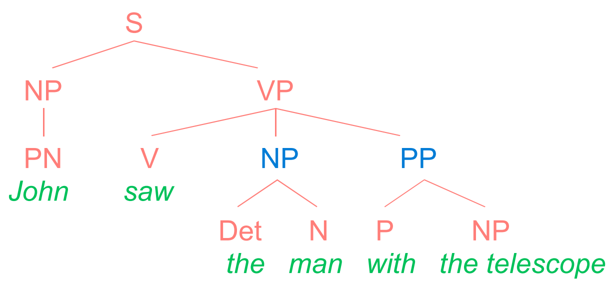
Prepositional Phase Attachment:



First Strategy: Minimal Attachment

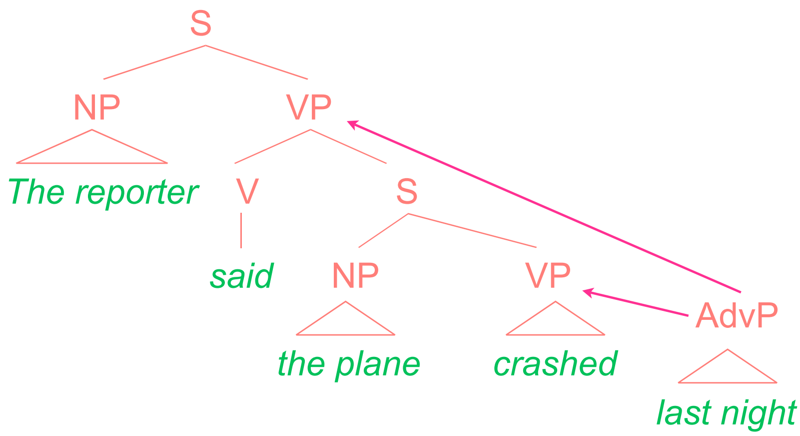
Minimal Attachment: Adopt the analysis which requires postulating the fewest nodes





Second Strategy: Late Closure

Late Closure: Attach material into the most recently constructed phrase marker



Summary of Frazier

Parsing preferences are guided by general principles:

- Serial structure building
- Reanalyze based on syntactic conflict
- Reanalyze based on low plausibility ("thematic fit")

Psychological assumptions:

- Modularity: only syntactic (not lexical, not semantic) information used for initial structure building
- Resources: emphasizes importance of memory limitations
- Processing strategies are universal, innate

Garden-Path Theory: Frazier (1978)

What **architecture** is assumed?

- Modular syntactic processor, with restricted lexical (category) and semantic knowledge

What **mechanisms** is used to construct interpretations?

- Incremental, serial parsing, with reanalysis

What **information** is used to determine preferred structure?

- General syntactic principles based on the current phrase stucture

Linking Hypothesis:

- Parse complexity and reanalysis cause increased RTs

Against linguistic modularity

Empirical evidence from on-line methods

- evidence for "immediate" (very early) interaction effects of animacy, frequency, plausibility, discourse context …
  - *The woman/patient sent the flowers was pleased*

Appropriate computational frameworks:

- symbolic constraint-satisfaction systems
- connectionist systems & competitive activation models

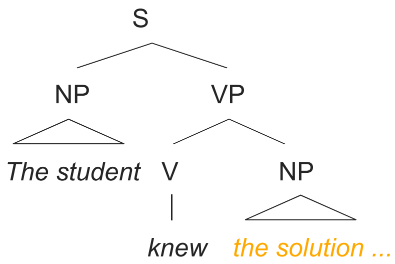
Homogenous/Integrative Linguistic Theory: HPSG

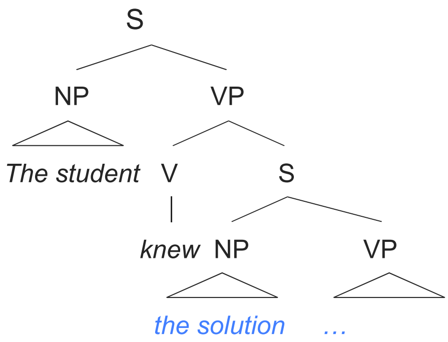
- multiple levels of representation within a unified formalism

NP/S Complement Ambiguity

The student knew the solution to the problem.

The student knew the solution was incorrect.





Grammar-Based Strategies

Not concerned with representation or ‘form’, but defined in terms of syntactic ‘content’

Strategies are modular, but ‘knowledge-based’

Motivation: strategies are derived from the purpose of the task, not e.g. computational efficiency

Closer competence-performance relationship

Pritchett (1992)

Rather than minimize complexity, maximize role assignment:

- Incrementally establish primary syntactic dependencies

Theta-Criterion: (GB theory, also in LFG + HPSG)

- Each argument must receive exactly one theta-role, and each theta role must be assigned to exactly one argument

Theta-Attachment:

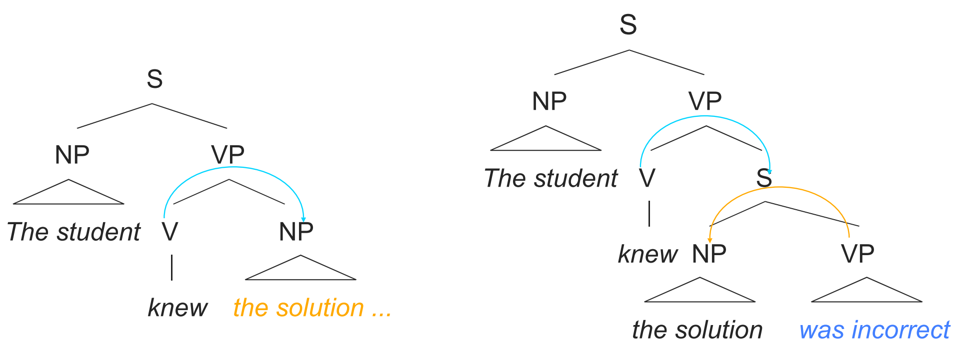
- Maximally satisfy the theta-criterion at every point during processing, given the maximal theta-grid of the verb

Theta Reanalysis Constraint:

- Reanalysis of a constituent out of its theta-domain results in a conscious garden-path effect

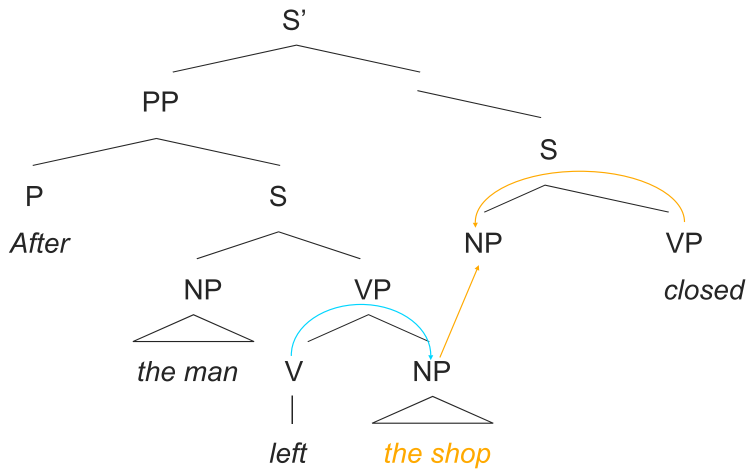
Theta-Reanalysis: Easy

Reanalysis to a position within the original theta-domain is easy.



Theta-Reanalysis: Difficult

Reanalysis to a position outside the original theta-domain is difficult.



Pritchett: Another example

"Without her contributions the orphanage closed"

- 'Without': a Prep with a single thematic role
- 'her':
  - an NP determiner of a yet unseen NP head, or
  - an Full NP complement (Pronoun), receives the role [Theta-attach]
- 'contributions':
  - head of a new NP, without a theta-role, or
  - build the larger NP with 'her', and receive the role [Theta-attach]

Well-known local ambiguities

NP/VP Attachment Ambiguity:

"The cop [saw [the burglar] [with the binoculars]]"

"The cop saw [the burglar [with the gun]]"

NP/S Complement Attachment Ambiguity:

"The athlete [realised [his goals]] last week"

"The athlete realised [[his goals] were unattainable]"

Clause-boundary Ambiguity:

"Since Jay always [jogs [a mile]] [the race doesn't seem very long]"

"Since Jay always jogs [[a mile] doesn't seem very long]"

Reduced Relative-Main Clause Ambiguity:

"[The woman [delivered the junkmail on Thursdays]]"

"[[The woman [delivered the junkmail]] threw it away]"

Relative/Complement Clause Ambiguity:

"The doctor [told [the woman] [that he was in love with her]]"

"The doctor [told [the woman [that he was in love with]] [to leave]]"

Grammar-Based (cont'd)

Theta-Attachment: reliance on theta-grids means it's head driven

- O.k. for English, but not incremental for head-final languages
- Same problem for Abney (1989), and other head-driven models

Pritchett's Theory (1992)

What **architecture** is assumed?

- Modular lexico-syntactic processor with syntactic and thematic role features

What **mechanisms** is used to construct interpretations?

- Incremental, serial parsing, with reanalysis

What **information** is used to determine preferred structure?

- Grammar principles and thematic role information

Linking Hypothesis:

- TRC violation causes garden-path, reanalysis without TRC is relatively easy

Experience and non-syntactic constraints

The previous accounts focus on

- Syntactic (and lexico-syntactic) ambiguity
- Purely syntactic mechanisms for disambiguation
- Assume a modular parser, the "primacy" of syntax

Does our prior **experience** with language, determines our preferences for interpreting the sentences we hear?

- Tuning hypothesis: disambiguate structure based on how it has been most frequently disambiguated in the past.

Non-syntactic **constraints**: to what extent do semantics, intonation, and context influence our resolution of ambiguity?

Multiple constraints in ambiguity resolution

The doctor told the woman that ...

story

diet was unhealthy

he was in love with her husband

he was in love with to leave

story was was about to leave

Prosody: intonation can assist disambiguation

Lexical preference: *that* = {Comp, Det, RelPro}

Subcat: *told* = { [_ NP NP] [_ NP S] [_ NP S'] [_ NP Inf] }

Semantics: Referential context, plausibility

- **Reference** may determine "argument attach" over "modifier attach"
- **Plausibility** of *story* versus *diet* as indirect object

Probabilistic Theories of Processing

Task of comprehension: recover the correct interpretation

Goal: Determine the most likely analysis for a given input:

$$\arg\max_i P(s_i) \text{ for all } s_i \in S$$

P can hide a multitude of sins:

- P corresponds to the degree of belief in an interpretation
- Influenced by recent utterances, experience, context

Implementation:

- P is determined by frequencies in corpora or completions
- To compare probabilities (of the Si), assume parallelism

Implementation

Interpretation of probabilities

- Likelihood of structure occurring, P can be determined by frequencies in corpora or human completions

Estimation of probabilities

- Infinite structural possibilities = sparse data
- Associate probabilities with grammar (finite): e.g. PCFGs

What mechanisms are required:

- Incremental structure building and estimation of probabilities
- Comparison of probabilities entails parallelism

Probabilistic Grammars

Context-free rules annotated with probabilities

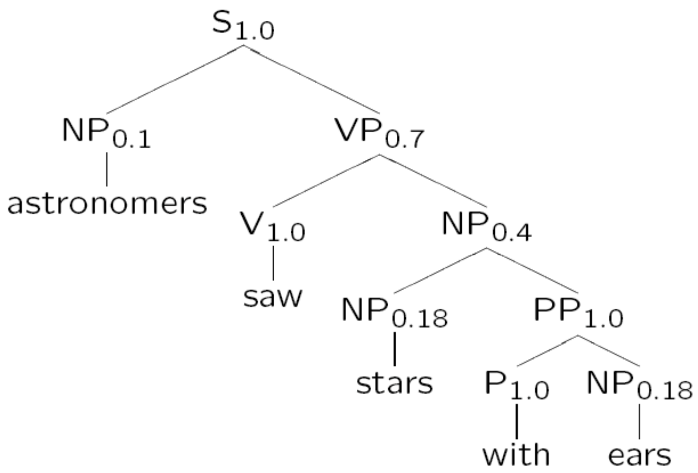
- Probabilities of all rules with the same LHS sum to one;
- Probability of a parse is the product of the probabilities of all rules applied in the parse.

Example (Manning and Schütze 1999)

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → VP NP	0.7	NP → ears	0.18
VP → VP NP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

Parse Ranking

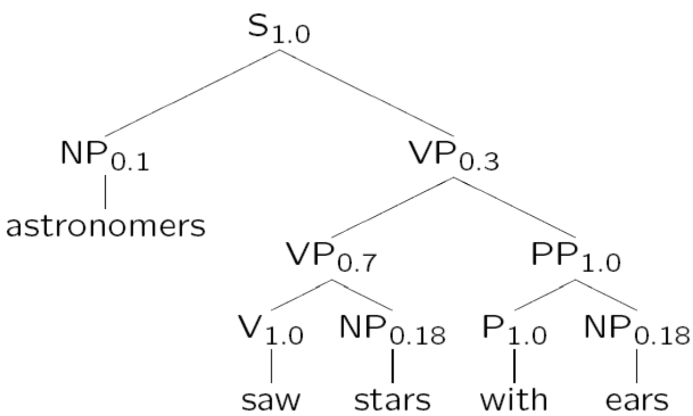
t_1:



$$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$

Parse Ranking

t_2:



$$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

Jurafsky (1996)

Probabilistic model of lexical and syntactic disambiguation

- exploits concepts from computational linguistics:
  - PCFGs, Bayesian modeling frame probabilities.

Overview of issues:

- data to be modeled: frame preferences, garden paths;
- architecture: serial, parallel, limited parallel;
- probabilistic CFGs, frame probabilities;
- examples for frame preferences, garden paths;
- comparison with other models; problems and issues.

Frame Preferences

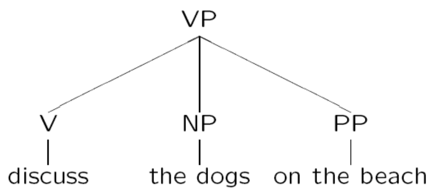
The women discussed the dogs on the beach.

- t1. The women discussed them (the dogs) while on the beach. (10%)
- t2. The women discussed the dogs which were on the beach. (90%)

$p(\text{discuss}, \langle \text{NP PP} \rangle) = 0.24$

VP → V NP XP 0.15

t_1:

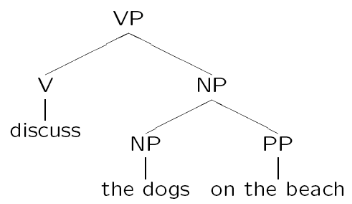


$p(t_1) = 0.15 \times 0.24 = 0.036$ (dispreferred)

$p(\text{discuss}, \langle \text{NP} \rangle) = 0.76$

VP → V NP 0.39
NP → NP XP 0.14

t_2:



$p(t_2) = 0.76 \times 0.39 \times 0.14 = 0.041$ (preferred)

Frame Preferences

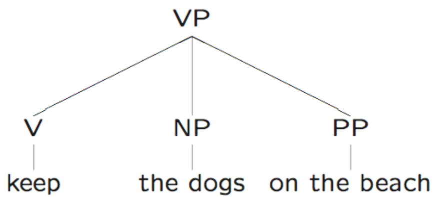
(2) The women kept the dogs on the beach.

- a. The women kept the dogs which were on the beach.
- b. **The women discussed them (the dogs) while on the beach.**

$p(\text{keep}, \langle \text{NP XP[pred +]} \rangle) = 0.81$

VP → V NP XP   0.15

t_1:

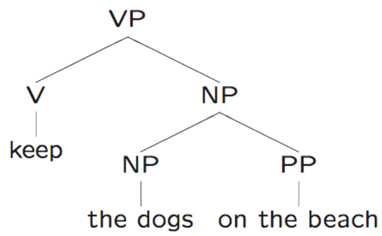


$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

$p(\text{keep}, \langle \text{NP} \rangle) = 0.19$

VP → V NP   0.39
NP → NP XP   0.14

t_2:



$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)

Modeling Garden Paths

The reduced relative clause often cause irrecoverable difficulty, but nor always:

- The horse raced past the barn fell (irrecoverable)
- The bird found died (recoverable)

We can use probabilities to distinguish the two cases, in a way a purely structural account (Frazier, or Pritchett) cannot.

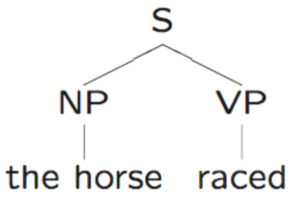
Assume a bounded, parallel parser …

- The parse with the highest probability is preferred
- Only those parsers which are within some "beam" of the preferred parse are kept, others are discarded

The horse raced past the barn fell

$p(\text{race}, \langle \text{NP} \rangle) = 0.92$

t_1:

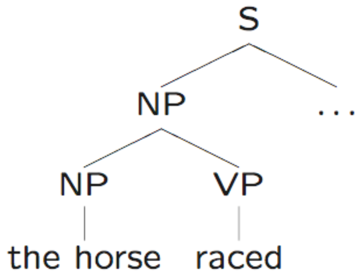


$p(t_1) = 0.92$ (preferred)

$p(\text{race}, \langle \text{NP NP} \rangle) = 0.08$

NP → NP XP 0.14

t_2:

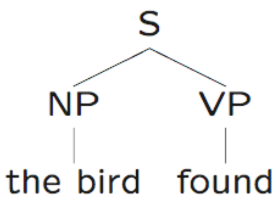


$p(t_1) = 0.0112$ (dispreferred)

The bird found died

$p(\text{find}, \langle \text{NP} \rangle) = 0.38$

t_1:

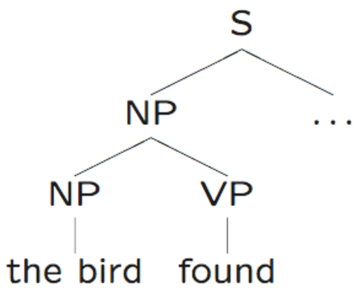


$p(t_1) = 0.38$ (preferred)

$p(\text{find}, \langle \text{NP NP} \rangle) = 0.62$

NP → NP XP 0.14

t_2:



$p(t_1) = 0.0868$ (dispreferred)

The Jurafsky Model

Setting the beam width:

- "The horse raced past the barn fell" 82:1
- "The bird found died" 4:1

Jurafsky assumes a garden path occurs (I.e. a parse is pruned) if its probability ratio with the best parse is greater than 5:1

Open issues:

- Where do we get the probabilities?
- Does the model work for other languages?

Garden-Path Theory: Jurafsky (1996)

What **architecture** is assumed?

- Modular lexico-syntactic processor with lexical (category and subcategory), no semantic knowledge

What **mechanisms** is used to construct interpretations?

- Incremental, bounded parallel parsing, with reranking

What **information** is used to determine preferred structure?

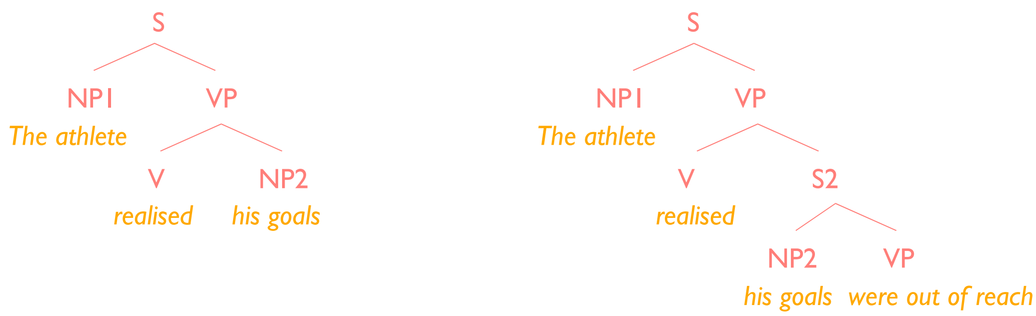
- Lexical and structural probabilities

Linking Hypothesis:

- Parse reranking causes increased RTs, if correct parse has been eliminated, predict a garden-path

A Problem for Likelihood?

NP/S Complement Ambiguity: *The athlete realised his goals ...*



Evidence for object attachment: (Pickering, Traxler & Crocker 2000)

- Despite S-comp bias of verb, NP is attached as D-object
- Ideal likelihood model and Jurafsky predict the opposite
- realised is initially tagged at S-comp, but the simpler DO analysis is then given higher probability, when NP is found