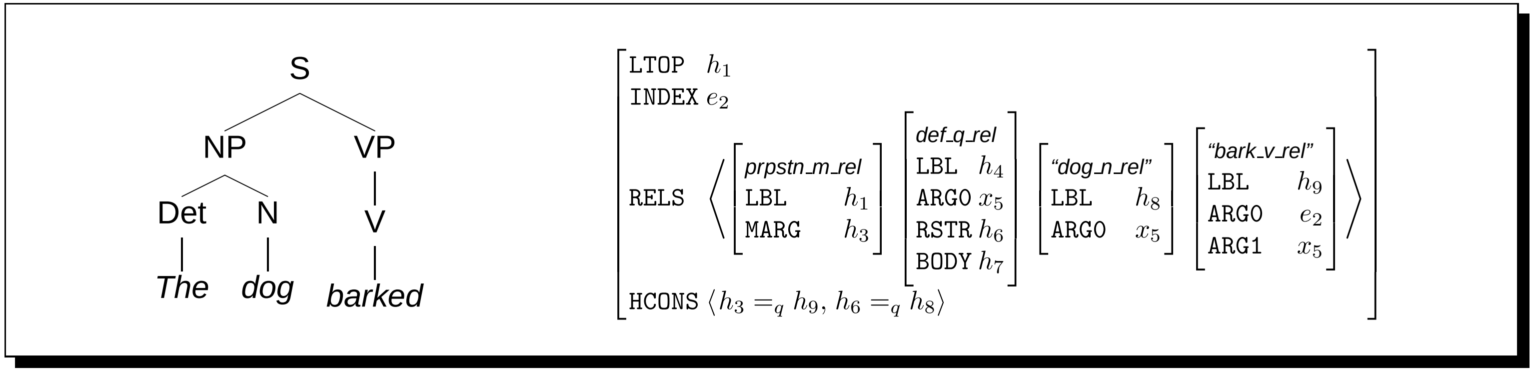


Language Technology Group (Statistical Disambiguation)

$$P(S \longrightarrow NP\ VP) = 1.0;\ \ P(NP \longrightarrow Det\ N) = 0.6$$

Stephan Oepen

Universitetet i Oslo

`oe@ifi.uio.no`

Ambiguity Resolution Remains a (Major) Challenge

The Problem

- With broad-coverage grammars, even moderately complex sentences typically have multiple analyses (dozens, sometimes tens of thousands);
- unlike in grammar writing, exhaustive parsing is useless for applications;
- identifying the ‘right’ (intended) analysis is an ‘AI-complete’ problem;
- inclusion of (non-grammatical) sortal constraints is generally undesirable.

Typical Approaches

- Design and use statistical models to select among competing analyses;
- for string s, some analyses t_i are more or less likely: maximize $P(t_i|s)$;

$\rightarrow$ Probabilistic Context Free Grammar (PCFG) is a CFG plus probabilities.

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

But it must be recognized that the notion 'probability' of a sentence is an entirely useless one, under any known interpretation of this term. (Noam Chomsky, 1969)

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

But it must be recognized that the notion 'probability' of a sentence is an entirely useless one, under any known interpretation of this term. (Noam Chomsky, 1969)

Every time I fire a linguist, system performance improves. (Fredrick Jelinek, 1980s)

Assigning Probabilities to Parse Trees

Treebanks

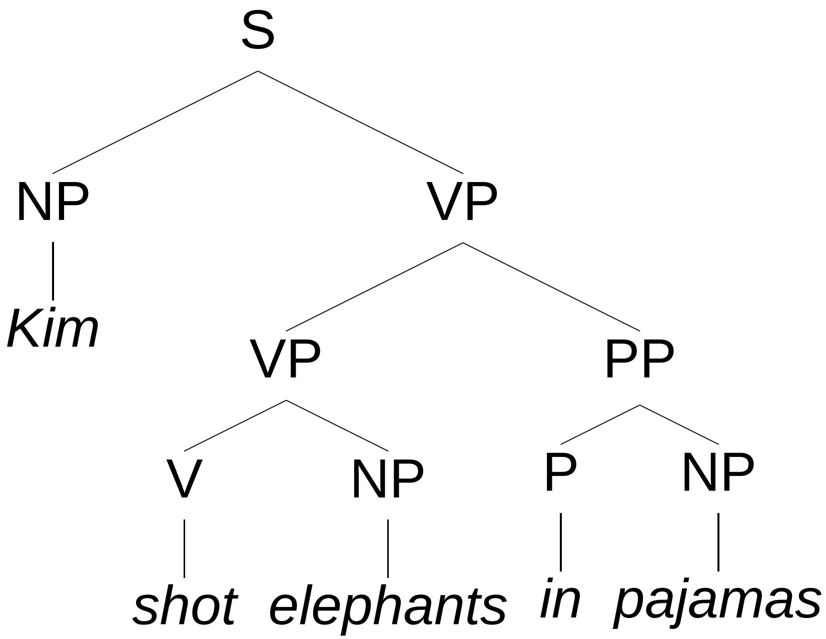
- For probability estimation, we need training data: ‘correct’ trees;
- a *treebank* pairs a corpus of sentences with *gold-standard* trees;
- *annotation* adds linguistic structure (e.g. trees) to raw corpus text;
- Penn Treebank: one million words of WSJ, manually annotated.

Probability Model

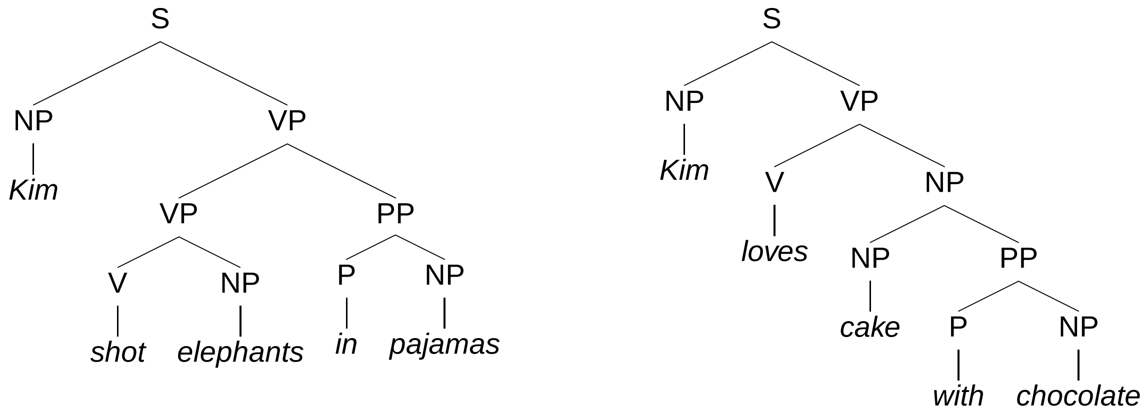
- A tree results from a sequence of rule applications (a derivation);
- joint probability: estimate *rule probabilities* and multiply (chain rule);
- assume that probability of each rule is independent from context.

A Quick Peek at the Penn Treebank

Consider a Practical Example



A (Simplified) PCFG Estimation Example



P(RHS|LHS)	**CFG Rule**
	S → NP VP
	VP → VP PP
	VP → V NP
	PP → P NP
	NP → NP PP
	VP → V

- Estimate rule probability from observed distribution;

→ conditional probabilities:

$$P(\text{RHS}|\text{LHS}) = \frac{C(\text{LHS}, \text{RHS})}{C(\text{LHS})}$$

Formally: Probabilistic Context-Free Grammars

- Formally, a *context-free grammar* (CFG) is a quadruple: $\langle C, \Sigma, P, S \rangle$

...

- P is a set of category rewrite rules (aka *productions*), each with a conditional probability P(RHS|LHS), e.g.

...

NP $\rightarrow$ Kim [0.6]
NP $\rightarrow$ snow [0.4]

...

- for each rule '$\alpha \rightarrow \beta_1, \beta_2, ..., \beta_n$' $\in P$: $\alpha \in C$ and $\beta_i \in C \cup \Sigma$; $1 \leq i \leq n$;

...

- for each $\alpha \in C$, the probabilities of all rules R '$\alpha \rightarrow$ ...' must sum to 1.

Parse Selection: The Maximum Entropy School

Conditional Parse Selection

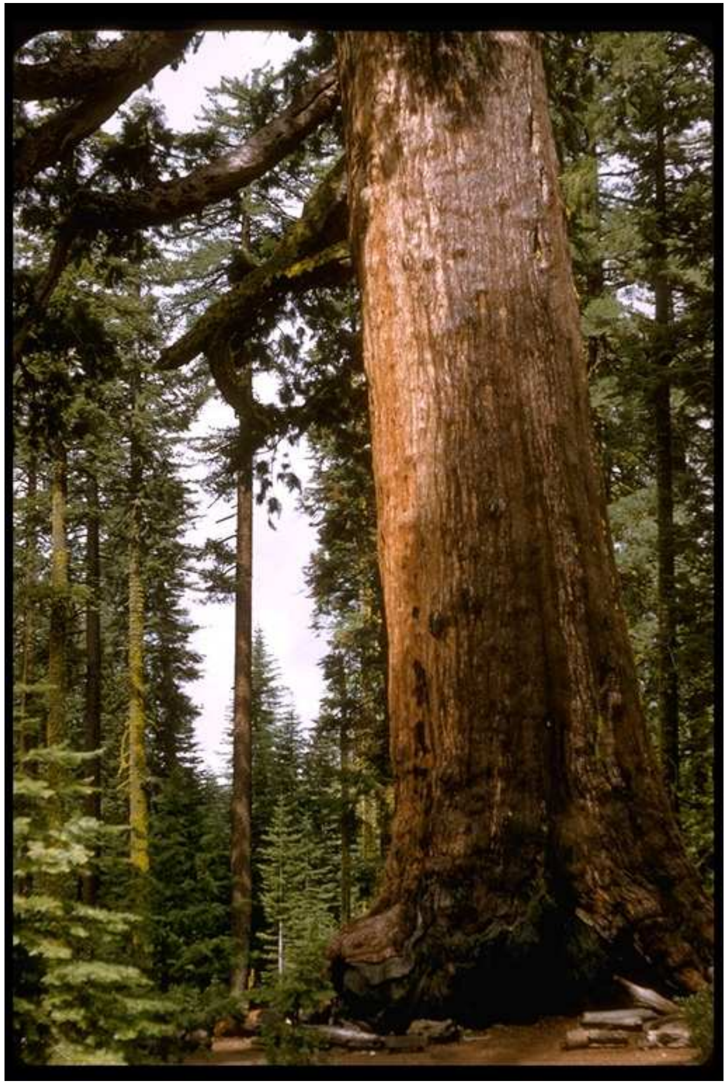
- Local independence assumption is not true for unification grammars;
- PCFG unable to 'learn' from negative data, e.g. dis-preferred parses;

$\rightarrow$ *conditional* model: given some context, sample properties of events.

Conditional Parse Selection

Given a sentence s and a set of trees $\{t_1 \ldots t_n\}$ assigned to s by some grammar, find the tree t_i that maximizes $p(t_i|s)$. Assuming a set of features $\{f_1 \ldots f_m\}$ with corresponding weights $\{\lambda_1 \ldots \lambda_m\}$, the conditional probability for tree t_i is given by:

$$p(t_i|s) = \frac{\exp \Sigma_j \lambda_j f_j(t_i)}{\Sigma_{k=1 \ldots n} \exp \Sigma_j \lambda_j f_j(t_k)} \quad (1)$$



LinGO Redwoods

— A Rich and Dynamic Treebank for HPSG —

**Stephan Oepen, Daniel P. Flickinger,
Kristina Toutanova, Christopher D. Manning**

Center for the Study of Language and Information
Stanford University

oe@csli.stanford.edu

Why (Yet) Another (Type of) Treebank?

Requirements for Disambiguation

- **syntax vs. semantics** topicalization vs. attachment ambiguity;
- **granularity** adequate match to degree of granularity in grammar;
- **adaptability** map into various formats; semi-automated updates.

Existing Resources (PTB, SUSANNE, NeGra, PDT, et al.)

- **(primarily) mono-stratal** topological *or* tectogrammatical;
- **(relatively) shallow** limited syntax, little or no semantics;
- **(mostly) static** (manual) ground truth annotation, no evolution.

LinGO Redwoods: a Rich and Dynamic Treebank

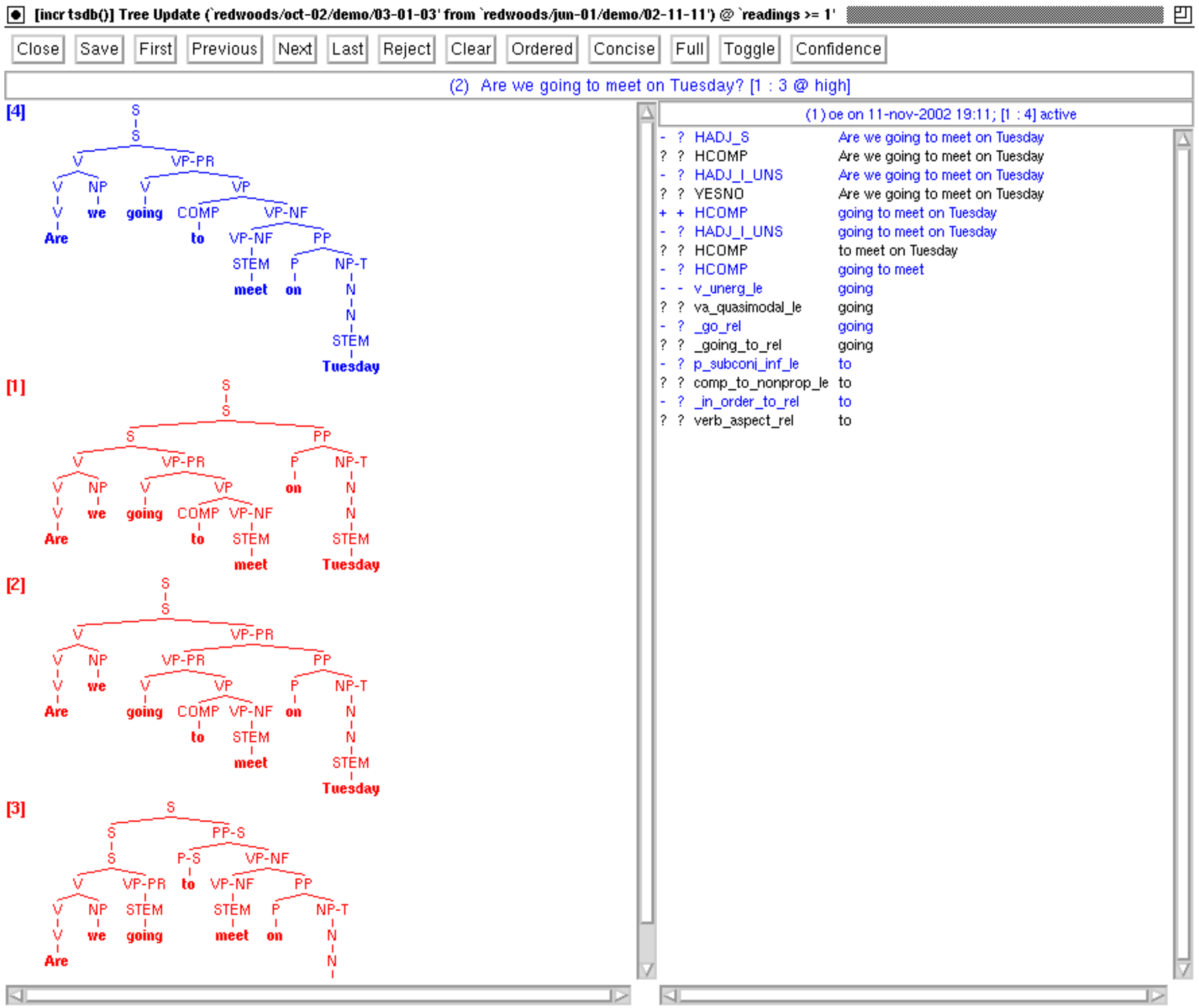
- Tie treebank development to existing broad-coverage grammar;
- hand-select (or reject) intended analyses from parsed corpus;
- [Carter, 1997]: annotation by basic discriminating properties;
- record annotator decisions (and entailment) as first-class data;
- provide toolkits for dynamic mappings into various formats;
- semi-automatically update treebank as the grammar evolves;
- integrate treebank maintenance with grammar regression testing.

Annotation: Basic Discriminating Properties

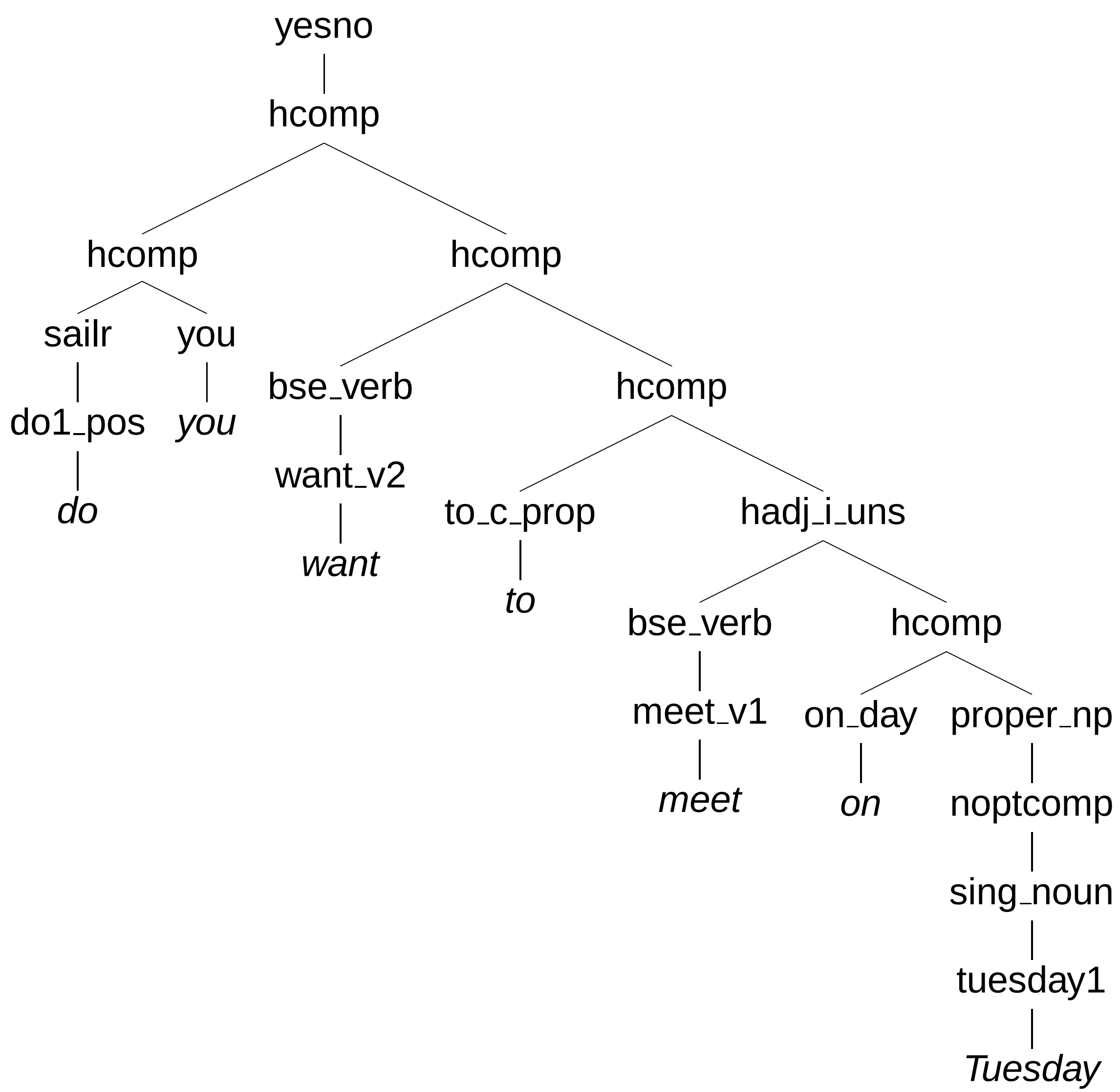
- Extract minimal set of *basic discriminants* from set of HPSG analyses;
- typically easy to judge, need little expert knowledge about grammar;
- allow quick navigation through parse forest and incremental reduction;
- *constituents*   use of particular construction over substring of input;
- *lexical items*   use of particular lexical entry for input token;
- *labeling*   assignment of particular abbreviatory label to a constituent;
- *semantics*   appearance of particular key relation on constituent;
- Stanford undergraduate annotates some 2000 sentences per week.

- Regularly propagate discriminants into new version of parsed corpus;

Redwoods Treebanking: A Quick Test Drive

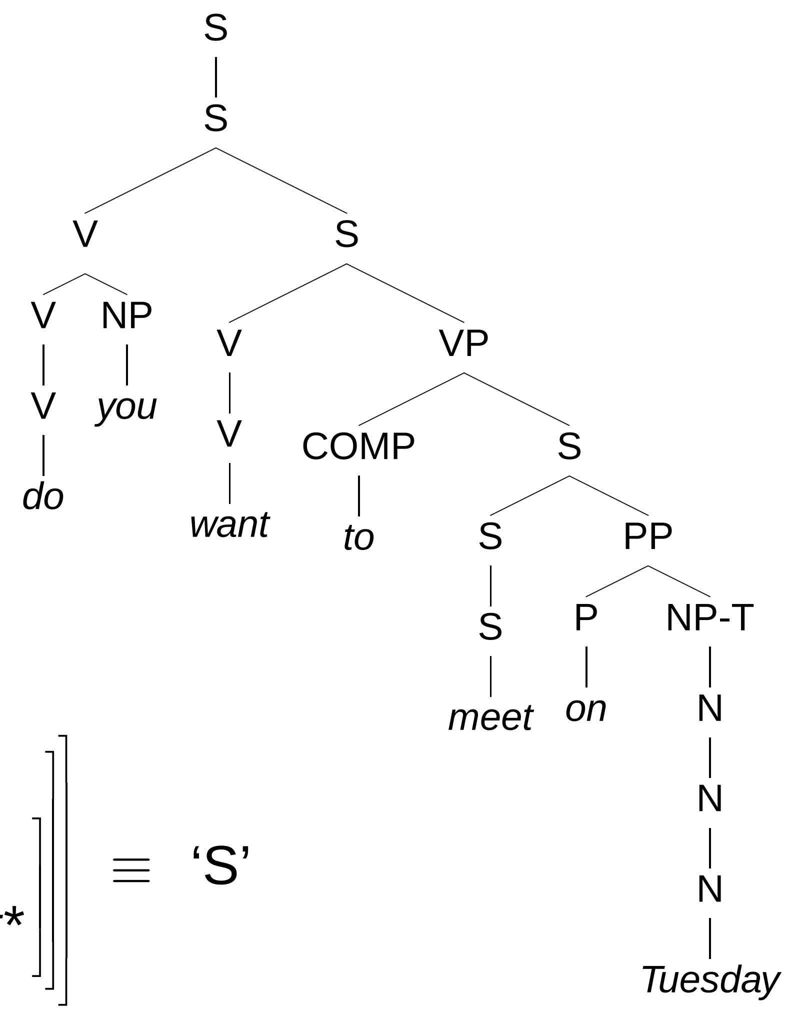


Redwoods Representations: Native Encoding



Derived Encodings: Labeled Phrase Structure Trees

- reconstruct full HPSG analysis from derivation tree;
- match underspecified feature structure ‘templates’ against each node;
- optionally, collapse or suppress nodes.



$$\textit{label}\left[\texttt{SYNSEM.LOCAL.CAT}\left[\begin{array}{ll}\texttt{HEAD} & \textit{verbal}\\ \texttt{VAL} & \left[\begin{array}{ll}\texttt{SUBJ} & \langle\rangle \\ \texttt{COMPS} & \textit{*olist*}\end{array}\right]\end{array}\right]\right] \equiv \text{‘S’}$$

Derived Encodings: Elementary Dependencies

- Reconstruct full HPSG analysis, compute MRS meaning representation;
- extract basic predicate – argument structure with uninterpreted roles;

→ labeled dependency graph fragments with (primarily) lexical relations.

```
e2:{
    _1:int_m[MARG _2:prpstn_m]
    _2:prpstn_m[MARG e2:_want_v_1]
    e2:_want_v_1[ARG1 x6:pron, ARG2 _3:prpstn_m]
    _3:prpstn_m[MARG e14:_meet_v_1]
    e14:_meet_v_1[ARG1 x6:pron]
    e15:_on_p_temp[ARG1 e14:_meet_v_1, ARG2 x16:dofw(tue)]
   }
```

Redwoods Development Status: 3rd Growth

	all parses			active $= 0$			active $= 1$			active > 1		
	$\sharp$	$\|$	$\times$	$\sharp$	$\|$	$\times$	$\sharp$	$\|$	$\times$	$\sharp$	$\|$	$\times$
VM_6	2706	7·7	46·7	216	9·4	63·5	2482	8·3	43·5	6	15·8	757·8
VM_{13}	2279	8·5	61·9	248	10·8	80·5	2029	8·7	59·5	2	15·5	198·0
VM_{31}	1967	6·2	27·9	216	10·1	95·9	1746	7·5	30·8	5	8·4	20·8
VM_{32}	699	7·5	53·2	15	11·8	57·7	684	8·4	53·2	0	0·0	0·0
Total	**7651**	**7·5**	**47·0**	**695**	**10·2**	**79·5**	**6941**	**8·2**	**45·9**	**13**	**12·9**	**388·2**

- 5th Growth release planned October 2004: up to 16,000 sentences;
- inclusion of ‘fragment’ utterances for VerbMobil: extra ambiguity;
- addition of ecommerce customer email corpus: 6,000 utterances.

Redwoods Applications: Parse Disambiguation

- Manning & Toutanova (Stanford): generative and conditional models;
- Baldridge & Osborne (Edinburgh): active learning and co-training;
- Fujita, Bond, et al. (NTT): semantics and ontologies in parse selection;
- feature selection: phrase structure, morpho-syntax, dependencies;
- ten-fold cross validation: score against annotated gold standard;
- preliminary results: 80^{+} % *exact match* parse selection accuracy;
- on-line use in parser: n-best beam search guided by MaxEnt scores;
- preferably, full parse forest (polynomial) plus selective unpacking.

Conclusions — Background Material

- 'Deep' grammar-based processing requires adequate stochastic models;
- basic research needed on acquisition and application of stochastic models;
- no existing treebank resources with suitable granularity and flexibility;
- LinGO Redwoods treebank based on existing open-source technology;
- tied to broad-coverage HPSG grammar: advantages and disadvantages;
- rich in available information, dynamic in data extraction and evolution.

Grammar and Treebank available from: *http://redwoods.stanford.edu/*

Based on Research and Contributions of

Tim Baldwin, John Beavers, Ezra Callahan
Emily M. Bender, Kathryn Campbell-Kibler,
John Carroll, Ann Copestake,
Rob Malouf, Ivan A. Sag,
Stuart Shieber, Tom Wasow,
and others.