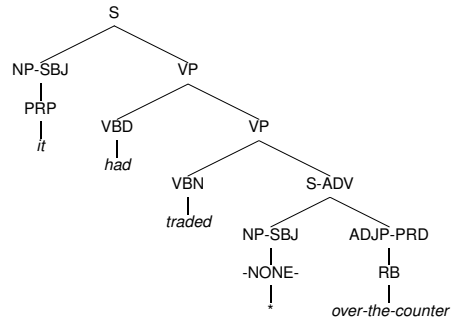


$$\begin{bmatrix} \texttt{LTOP} & h_1 \\ \texttt{INDEX} & e_2 \\ \texttt{RELS} & \left\langle \begin{bmatrix} \textit{prpstn\_m\_rel} \\ \texttt{LBL} & h_1 \\ \texttt{MARG} & h_3 \end{bmatrix} \begin{bmatrix} \textit{def\_q\_rel} \\ \texttt{LBL} & h_4 \\ \texttt{ARG0} & x_5 \\ \texttt{RSTR} & h_6 \\ \texttt{BODY} & h_7 \end{bmatrix} \begin{bmatrix} \textit{“dog\_n\_rel”} \\ \texttt{LBL} & h_8 \\ \texttt{ARG0} & x_5 \end{bmatrix} \begin{bmatrix} \textit{“bark\_v\_rel”} \\ \texttt{LBL} & h_9 \\ \texttt{ARG0} & e_2 \\ \texttt{ARG1} & x_5 \end{bmatrix} \right\rangle \\ \texttt{HCONS} & \langle h_3 =_q h_9, h_6 =_q h_8 \rangle \end{bmatrix}$$

Treebanks vs. Linguistic Theory

— Anyone Need Grammars? —

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

oe@csli.stanford.edu

A Wake-Up Call to Our Community

Every time I fire a linguist,
system performance goes up.

[Fred Jelinek, 1980s]

A Wake-Up Call to Our Community

Every time I fire a linguist,
system performance goes up.

[Fred Jelinek, 1980s]

A Whole Generation of Traumatized Linguists

- Streams of fashion: analytical vs. empirical, linguistic vs. data-driven;
- (perceived) paradigm shift in the 1990s: discontinue 'deep' processing;
- Jelinek eventually turned off the lights — LFG & HPSG groups stable;

→ keep focus: 'deep' linguistic approaches required for long-term success.

A Few More (Manufactured) Quotes

To me, the ultimate goal of our new field of Computational Linguistics is the automated translation of human language.

(Martin Kay, 1960s)

A Few More (Manufactured) Quotes

To me, the ultimate goal of our new field of Computational Linguistics is the automated translation of human language.

(Martin Kay, 1960s)

Computational linguists believe that by throwing more cycles and more raw text into their statistical black box, they can dispense with linguists altogether. [...] Anyone who cannot at least use the [statistical] terminology persuasively risks being mistaken for kitchen help at the ACL banquet. (Steven Abney, 1996)

A Few More (Manufactured) Quotes

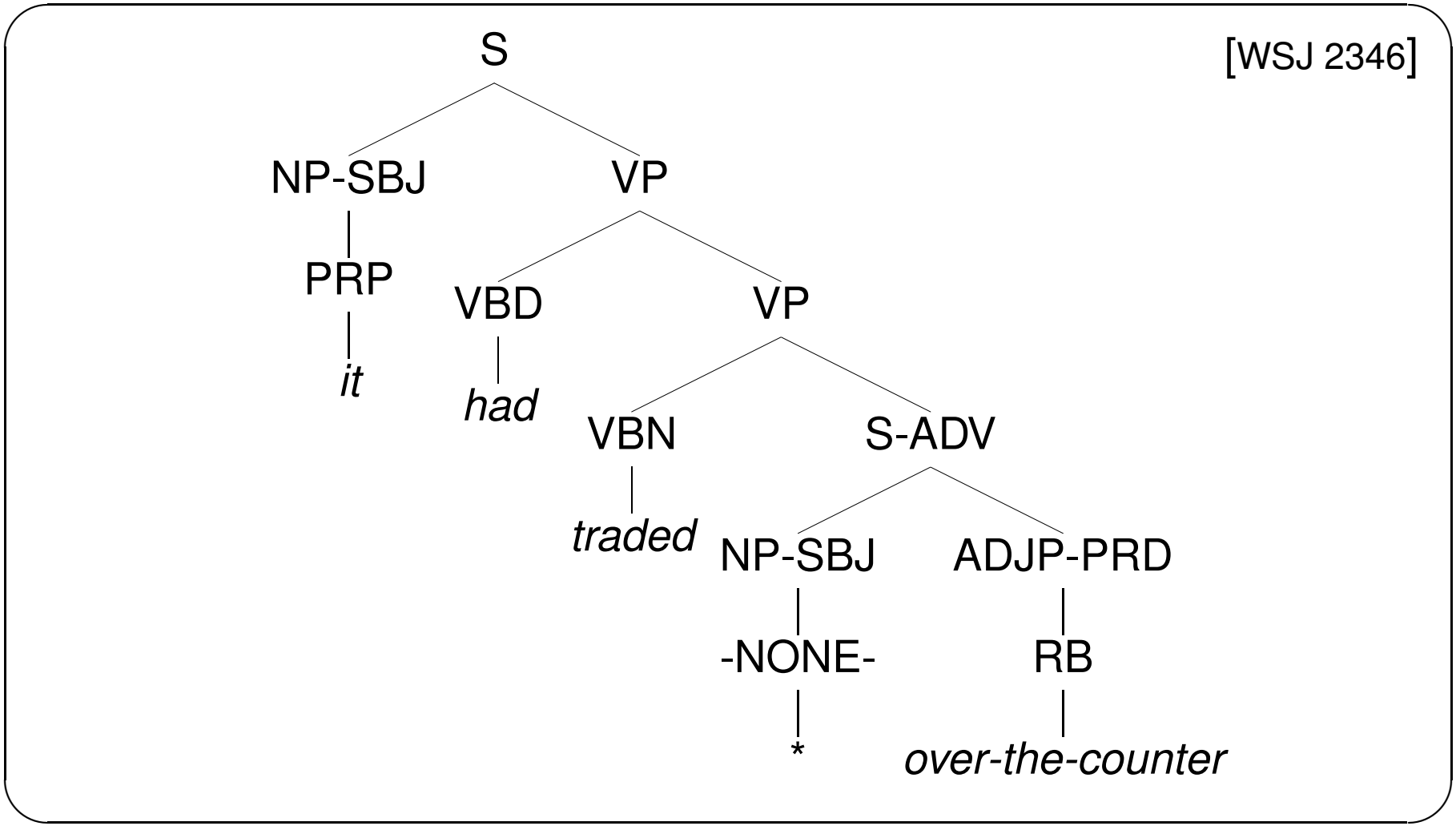
To me, the ultimate goal of our new field of Computational Linguistics is the automated translation of human language.

(Martin Kay, 1960s)

Computational linguists believe that by throwing more cycles and more raw text into their statistical black box, they can dispense with linguists altogether. [...] Anyone who cannot at least use the [statistical] terminology persuasively risks being mistaken for kitchen help at the ACL banquet. (Steven Abney, 1996)

We should probably all spend more time on the linguistic annotation of actual data rather than on writing grammar rules, based primarily on introspection. (Erhard Hinrichs, 1990s)

Linguistic Theory Can be Hard to Find



A Stretch of Our Imagination

The Ultimate Grammar

- Full linguistic coverage on arbitrary data, cross-domain and -genre;
- adequate grammatical analyses in all cases; inclusion of semantics;
- fully declarative; same grammar for both parsing and generation;
- high-efficiency processing tools: (minimally) real-time performance.

The Final Treebank

- Representative data for ‘all’ of the language, domains, and genres;
- full annotation with (at least) syntactic and semantic information;
- utterly coherent, free of errors, fully documented, freely available.

Do We Have That Much Imagination?

Theory and Grammar Building

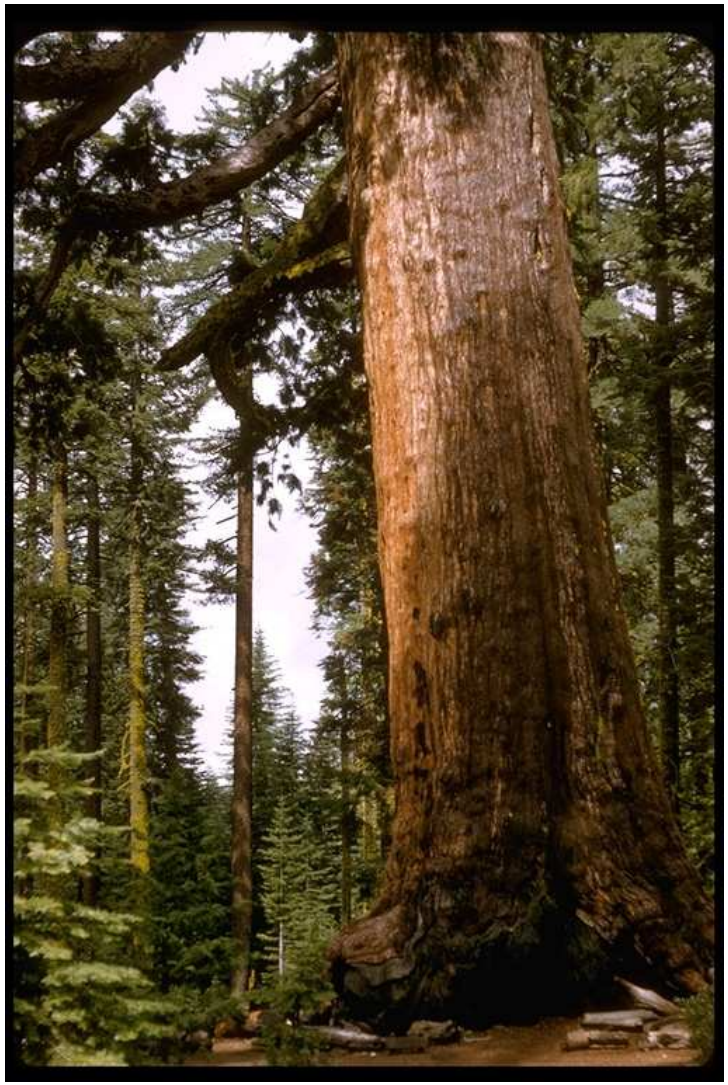
- No generally accepted linguistic theory, broad-coverage analyses;
- long, tedious, error-prone engineering process; rather few experts.

Corpora Creation

- No generally agreed upon standard of depth and type of information;
- lack of disk space; long, tedious, error-prone annotation process.

Which of the two dreams is more likely to come true (soon)?

Which of the two resources will bring us closer to the ultimate goal?



LinGO Redwoods

— A Rich and Dynamic Treebank for HPSG —

Stephan Oepen, Dan Flickinger,
Kristina Toutanova, and
Christopher D. Manning

Center for the Study of Language and Information
Stanford University

oe@csli.stanford.edu

A Candidate: LinGO English Resource Grammar

Development Background (1993 – present)

- General-purpose, wide-coverage, computational English grammar;
- mainly Dan Flickinger, with Rob Malouf, Emily M. Bender, Jeff Smith;
- supported in multiple HPSG processing environments (LKB & PET).

Design

- HPSG [Pollard & Sag 1994]: constraint-based, strongly lexicalized;
- MRS [Copestake et al., 1999]: flat, event-based, underspecified;
- type hierarchies defining principles, lexical classes, constructions;
- strict grammaticality assumption: generator using same grammar.

LinGO ERG: Coverage and Size

Linguistic Coverage

- 85 % of 12,000 transcribed dialogue turns from VerbMobil domains;
- 80^{+} % of customer emails in financial and ecommerce domains;
- both fairly short utterances: average 9 words, ranging from $1-40$;
- 80 % of phenomena-based examples in Hewlett Packard test suite.;
- more recently, 95 % on excerpts from tourism brochures (13 words).

Size of Grammar (as of October 2003)

- some 2,600 types for fundamentals, lexicon, rules, and sematics;
- 11,152 lexical entry stems (around 2,500 verbs and 3,100 nouns);
- 27 lexical (15 inflectional) rules and 96 phrase structure schemata.

Sample Data (LOGON Domain)
Analyzed by LinGO English Grammar

1 *Be considerate of game, farm animals and other hikers.*

109 *Kjeragveggen has interested climbers since the 1970s.*

304 *But there are things to do for those with knickers and anoraks too.*

39 *Follow the road past NUTEC and continue up along Kvarvenveien, past the recreation area.*

248 *The first part of the trip goes with the Hurtigruta to Torvik, with a bicycle ride at night into the sunrise out to Runde, and a hike to Norway's southernmost bird mountain.*

326 *If there is one thing Swedes are concerned with, it is preparing delicious dishes.*

Grammatical Coverage on Tourism Excerpts

‘lingo/08-nov-03/hike/03-11-14/pet’ Coverage Profile						
Aggregate	**total items** ♯	**word string** ϕ	**lexical items** ϕ	**parser analyses** ϕ	**total results** ♯	**overall coverage** %
$35 \leq$ *i-length* < 40	1	35.00	109.00	2372.00	1	100.0
$30 \leq$ *i-length* < 35	2	32.50	109.00	1768.00	2	100.0
$25 \leq$ *i-length* < 30	7	26.71	100.57	1393.14	7	100.0
$20 \leq$ *i-length* < 25	28	21.68	78.36	931.93	28	100.0
$15 \leq$ *i-length* < 20	72	16.89	54.08	136.18	67	93.1
$10 \leq$ *i-length* < 15	119	11.77	39.85	35.87	113	95.0
$5 \leq$ *i-length* < 10	95	7.47	23.49	5.79	89	93.7
$0 \leq$ *i-length* < 5	6	4.00	7.67	1.33	6	100.0
Total	**330**	**12.86**	**42.85**	**177.17**	**313**	**94.8**

(generated by [incr tsdb()] at 14-nov-2003 (22:49 h))

Ambiguity Resolution Remains a (Major) Challenge

The Problem

- With broad-coverage grammars, even moderately complex sentences typically have multiple analyses (tens or hundreds, rarely 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.

'Current' State of Affairs

- Heuristic scoring rules applied to (classes of) lexical items and rules;
- LFG 'optimality' projection: accumulate quality marks and rank globally;
- embryonic work on probabilistic models for on- or off-line parse selection.

(At Least) Three Dimensions to the Problem

Unsupervised vs. Semi-Supervised vs. Supervised

- Costly construction of hand-disambiguated training data (Redwoods);
- un- and semi-supervised learning using CFGs and EM (CoLi & DFKI).

Source Grammar vs. (CF) Approximation

- Feature selection from broad range of syntactic and semantic properties;
- simpler (faster) models based on CF approximation; prune search space.

On-Line ('Dynamic') vs. Off-Line Disambiguation

- Exponential factor in scoring *all* analyses from full grammar after parsing;
- practical applications will likely need on-line models or phased scheme.

Why (Yet) Another (Type of) Treebank?

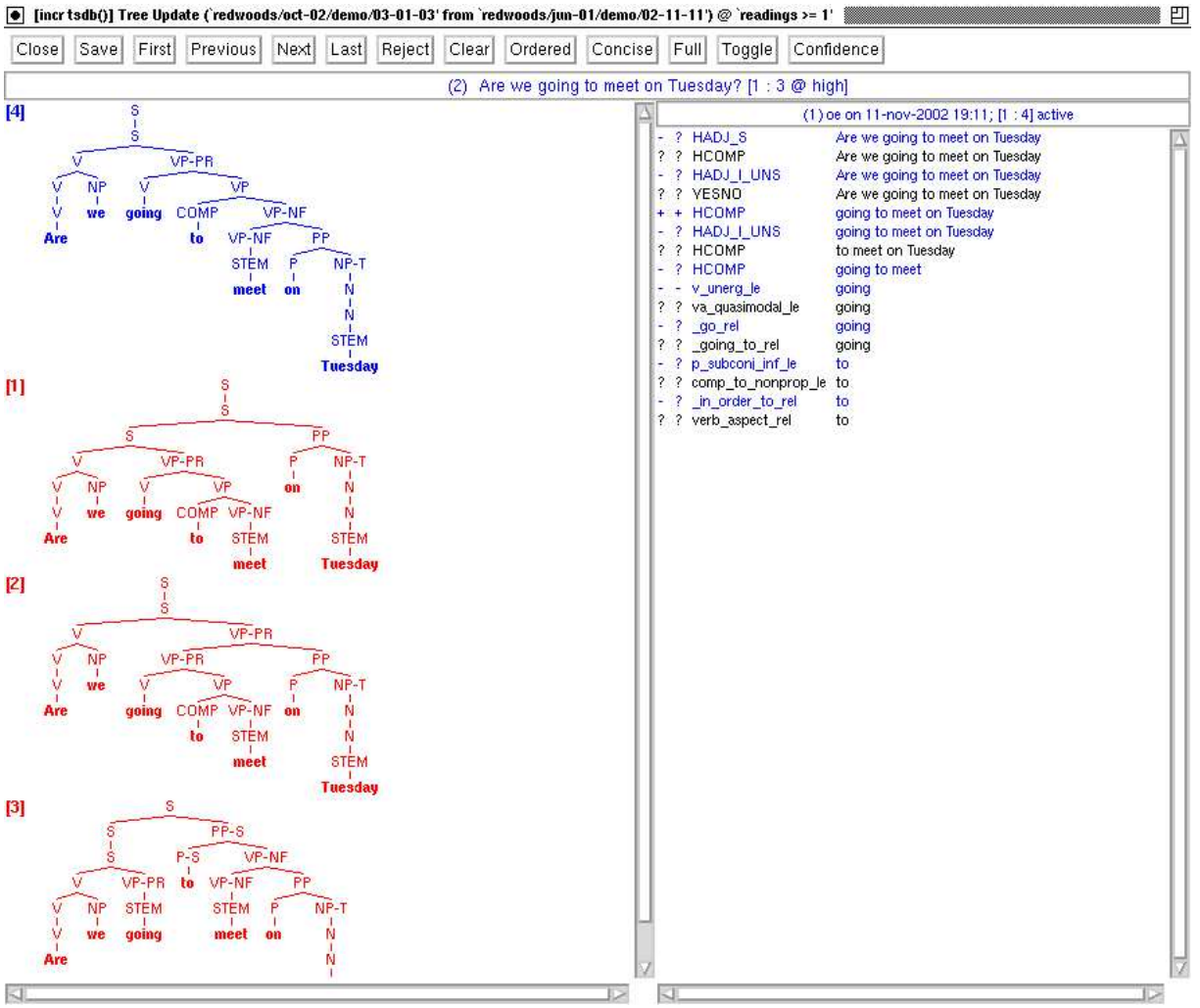
Requirements for Disambiguation

- **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 Quick Test Drive

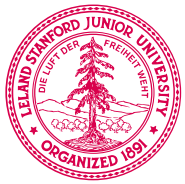


LinGO Redwoods: a Rich and Dynamic Treebank

- 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;
- integrate treebank maintenance with grammar regression testing.

Key Challenges

- Derivative of grammar: undergeneration results in gaps in treebank;
- grammar evolution gradually invalidates treebank; update procedures.



Annotation: Basic Discriminating Properties

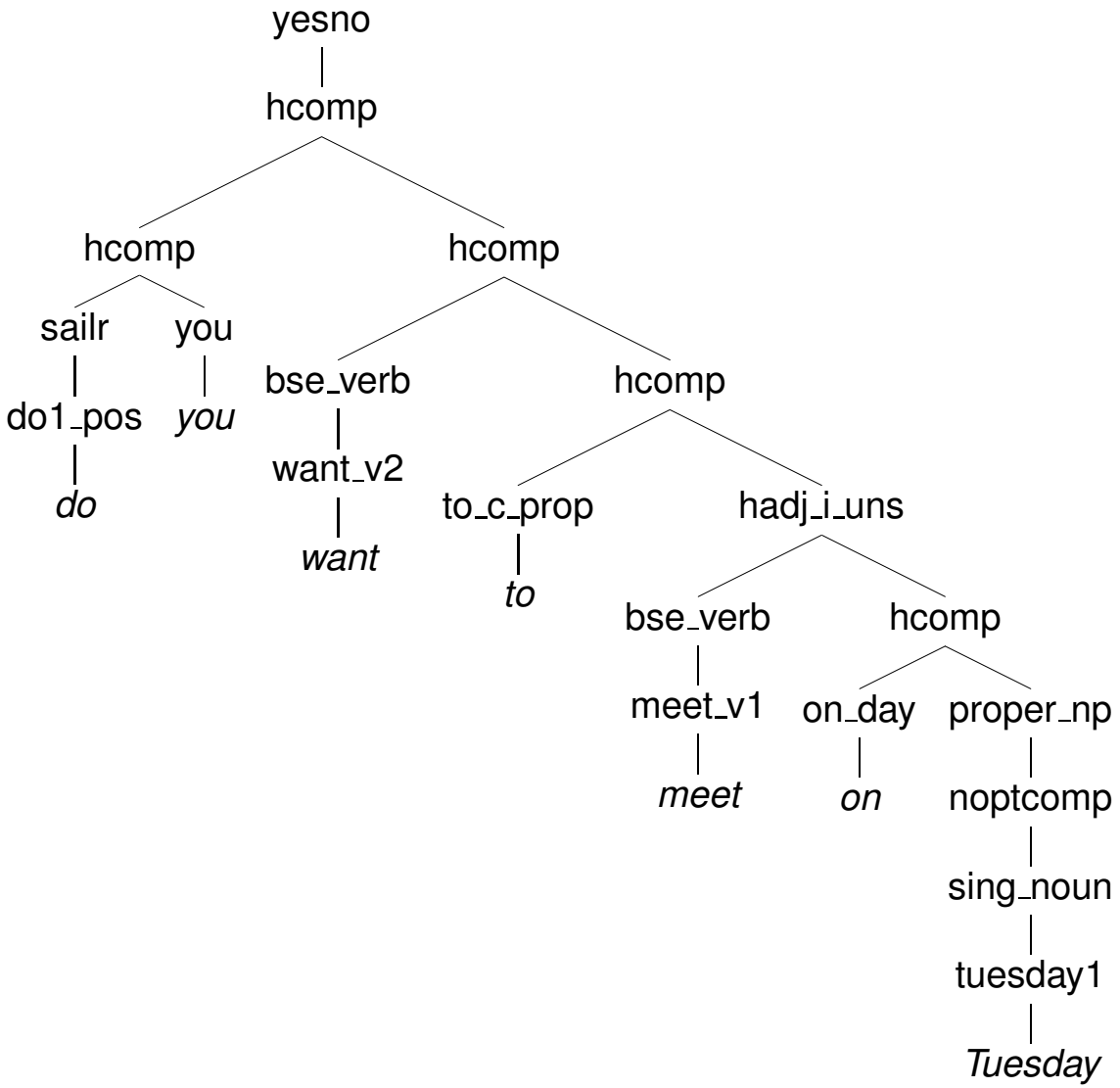
Key Notions

- Extract minimal set of *basic discriminants* from set of HPSG analyses;
- quick navigation through parse forest; easy to judge [Carter, 1997];
- constituents: use of particular construction over substring of input;
- lexical items: use of particular lexical entry for input token (a 'word');
- labeling: assignment of particular abbreviatory label to a constituent;
- semantics: appearance of particular key relation on constituent.

Preliminary Experience

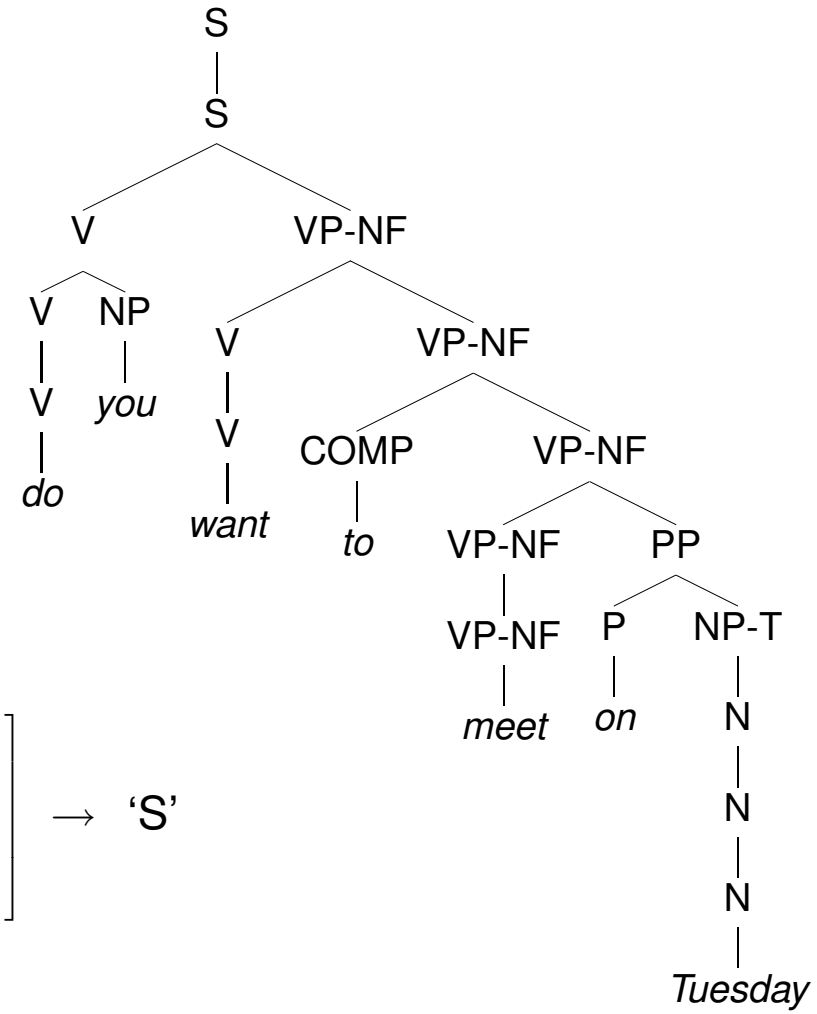
- Stanford undergraduate annotates some 2000 sentences per week.

Redwoods Representations: Native Encoding



Derived Encodings: Labeled Phrase Structure Trees

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



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

Derived Encodings: Elementary Dependencies

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

→ labeled dependency graph fragments of (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		
	♯	‖	×	♯	‖	×	♯	‖	×	♯	‖	×
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**

‘♯’ total number of items (sentences) in each aggregate;

‘‖’ average item length (number of input tokens in string);

‘×’ average structural ambiguity (number of analyses).

Redwoods Applications: Parse Disambiguation

- Manning & Toutanova (Stanford): generative and conditional models;
- Baldridge & Osborne (Edinburgh): active learning and co-training;
- restrict to Redwoods subset of fully disambiguated ambiguous items;
- 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;

$\rightarrow$ native encoding performs far better than labeled constituent trees.

Semi-Automatic Update Procedure

Bi-Weekly Internal Releases of Revised Grammar

- Regularly, with new grammar version, obtain updated parsed corpus;
- propagate annotator decisions (discriminants), primary and entailed.
- new ambiguity: distinctions added to the grammar, manual resolution;
- invalid or spurious discriminants: distinctions lost or reformulated;
- 'misleading' discriminants: theoretically possible but (highly) unlikely;
- inspection of mismatches provides diagnostic feedback to grammar;
- integration with grammar development cycle, minimize manual work.

LinGO ERG: June 2001 vs. October 2002

	jun-01	oct-02	△
appropriate features	148	149	−6% +7%
type hierarchy (excluding lexicon)	3,062	3,895	+27%
grammar rules (including lexical rules)	86	94	−11% +26%
lexical types (‘parts of speech’)	400	580	+45%
semantic relations (‘predicates’)	5,406	6,162	+14%
lexical entries	8,135	9,954	+22%
lines of source (excluding lexicon)	25,847	32,199	+25%

Semi-Automated Updates: It Actually Works

Aggregate	items	original in	original out	matches yes	matches no	update in	update out	update new	final in	final out
	♯	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ
new = 0	1421	1·1	23·6	8·1	8·5	1·0	13·9	0·0	1·0	13·9
new = 1	708	1·1	38·1	6·9	9·8	2·2	29·6	1·0	1·0	30·8
new ≥ 2	273	1·3	61·5	12·1	15·2	4·2	72·0	2·8	1·0	75·2
Total	**2402**	**1·1**	**32·2**	**8·2**	**9·6**	**1·8**	**25·1**	**0·6**	**1·0**	**25·9**
new = 0	2195	1·0	72·2	17·2	1·0	1·0	69·3	0·0	1·0	69·3
new = 1	73	1·0	31·9	11·7	1·4	2·2	116·0	1·0	1·0	117·3
new ≥ 2	20	1·0	192·6	13·3	0·8	16·7	297·5	2·9	1·0	313·2
Total	**2288**	**1·0**	**72·0**	**17·0**	**1·1**	**1·2**	**72·8**	**0·1**	**1·0**	**73·0**

Related Work

Non-Public Environments

- Related work at SRI Cambridge, (Xerox) PARC, and M$ Research;
- grammars, language corpora, and treebanks not publicly available;
- results published in some cases, generally difficult to reproduce.

Academic Environments

- [Dipper, 2000] LFG for German, 'transfer' into TiGer format;
- [Bouma et al., 2001] HPSG for Dutch, dependency structures only;
- [Simov et al., 2002] parallel treebanking and grammar writing;
- to our best knowledge, no existing *rich* and *dynamic* treebanks.

More (Authorized, for a Change) Quotes

Given recent theoretical, methodological, and technological advances, maybe more people should consider building grammars of long-term value rather than expending effort on mid-term stopgaps. (Dr. phil. Stephan Oepen, 2003)

More (Authorized, for a Change) Quotes

Given recent theoretical, methodological, and technological advances, maybe more people should consider building grammars of long-term value rather than expending effort on mid-term stopgaps. (Dr. phil. Stephan Oepen, 2003)

Treebanks can hardly be a goal in themselves, but serve an important purpose — as a means to an end. (Dr. phil. Stephan Oepen, 2003)

More (Authorized, for a Change) Quotes

Given recent theoretical, methodological, and technological advances, maybe more people should consider building grammars of long-term value rather than expending effort on mid-term stopgaps. (Dr. phil. Stephan Oepen, 2003)

Treebanks can hardly be a goal in themselves, but serve an important purpose — as a means to an end. (Dr. phil. Stephan Oepen, 2003)

For machine translation, I certainly see more long-term potential in statistical approaches, simply because these systems have the ability to learn. We will talk again in ten years. (Dr. ing. Thorsten Brants, 2003)

More (Authorized, for a Change) Quotes

Given recent theoretical, methodological, and technological advances, maybe more people should consider building grammars of long-term value rather than expending effort on mid-term stopgaps. (Dr. phil. Stephan Oepen, 2003)

Treebanks can hardly be a goal in themselves, but serve an important purpose — as a means to an end. (Dr. phil. Stephan Oepen, 2003)

For machine translation, I certainly see more long-term potential in statistical approaches, simply because these systems have the ability to learn. We will talk again in ten years. (Dr. ing. Thorsten Brants, 2003)

People who like grammars need grammars. (Abraham Lincoln, 1860s)

Conclusions — Outlook

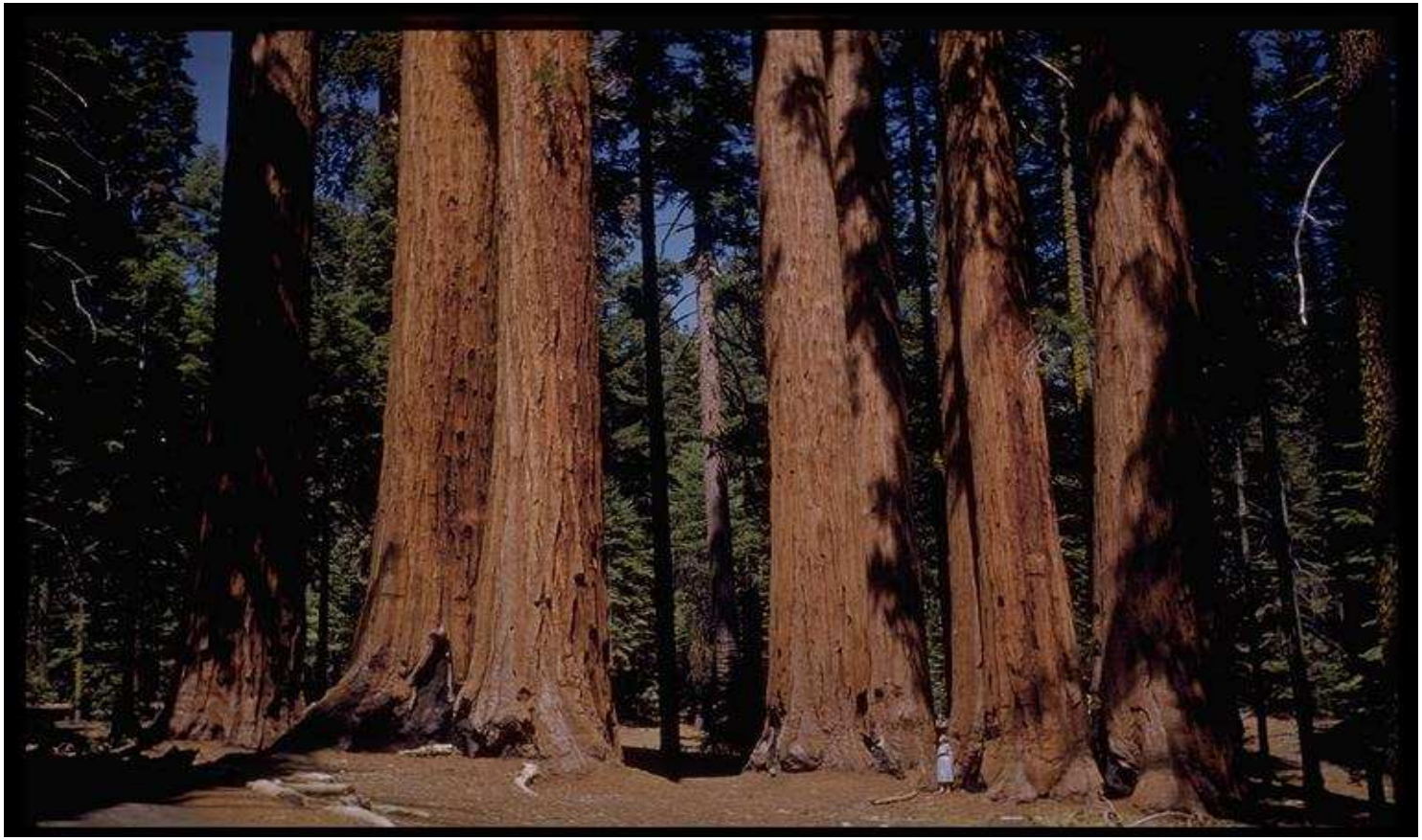
- ‘Deep’ grammar-based processing requires adequate stochastic models;
- no existing treebank resources with suitable granularity and flexibility;
- LinGO Redwoods treebank tied to broad-coverage HPSG implementation;

→ rich in available information, dynamic in data extraction and evolution.

More Recent Developments

- Annotation of some 3,000 customer emails from ecommerce domain;
- Japanese off-spring: *Hinoki* (NTT); 92 % coverage on dictionary definitions;
- approach and tools adapted to more ‘shallow’ frameworks: RASP & VISL.

Outlook: Go, Take a Stroll!



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,
Dan Flickinger, Rob Malouf, Chris Manning,
Ivan A. Sag, Stuart Shieber,
Kristina Toutanova, Tom Wasow,
and others.