Towards Holistic Testing

Grafting Treebank Maintenance into the
Grammar Revision Cycle

Stephan Oepen
Universitetet i Oslo
& CSLI Stanford
oe@csli.stanford.edu

Dan Flickinger
CSLI Stanford
danf@csli.stanford.edu

Francis Bond
NTT Communication Science Laboratory
bond@cslab.kecl.ntt.co.jp

Why Both a Grammar and a Treebank?

Ambiguity Management

- 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;

→ emerging work on stochastic parse selection requires training material.

Sustained Coverage

- Large-scale grammars are intricate: systematic regression testing;
- variety of constructions across different data sets; corpora and test suites;

→ need to identify and maintain *intended* analyses, trees, semantics, et al.

Our Grammar: 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 (Tourism Domain)
Analyzed by LinGO ERG

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

	total	**word**	**lexical**	**parser**	**total**	**overall**
Aggregate	**items**	**string**	**items**	**analyses**	**results**	**coverage**
	$\sharp$	ϕ	ϕ	ϕ	$\sharp$	$\%$
$35 \leq \textit{i-length} < 40$	1	35.00	109.00	2372.00	1	100.0
$30 \leq \textit{i-length} < 35$	2	32.50	109.00	1768.00	2	100.0
$25 \leq \textit{i-length} < 30$	7	26.71	100.57	1393.14	7	100.0
$20 \leq \textit{i-length} < 25$	28	21.68	78.36	931.93	28	100.0
$15 \leq \textit{i-length} < 20$	72	16.89	54.08	136.18	67	93.1
$10 \leq \textit{i-length} < 15$	119	11.77	39.85	35.87	113	95.0
$5 \leq \textit{i-length} < 10$	95	7.47	23.49	5.79	89	93.7
$0 \leq \textit{i-length} < 5$	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))

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.

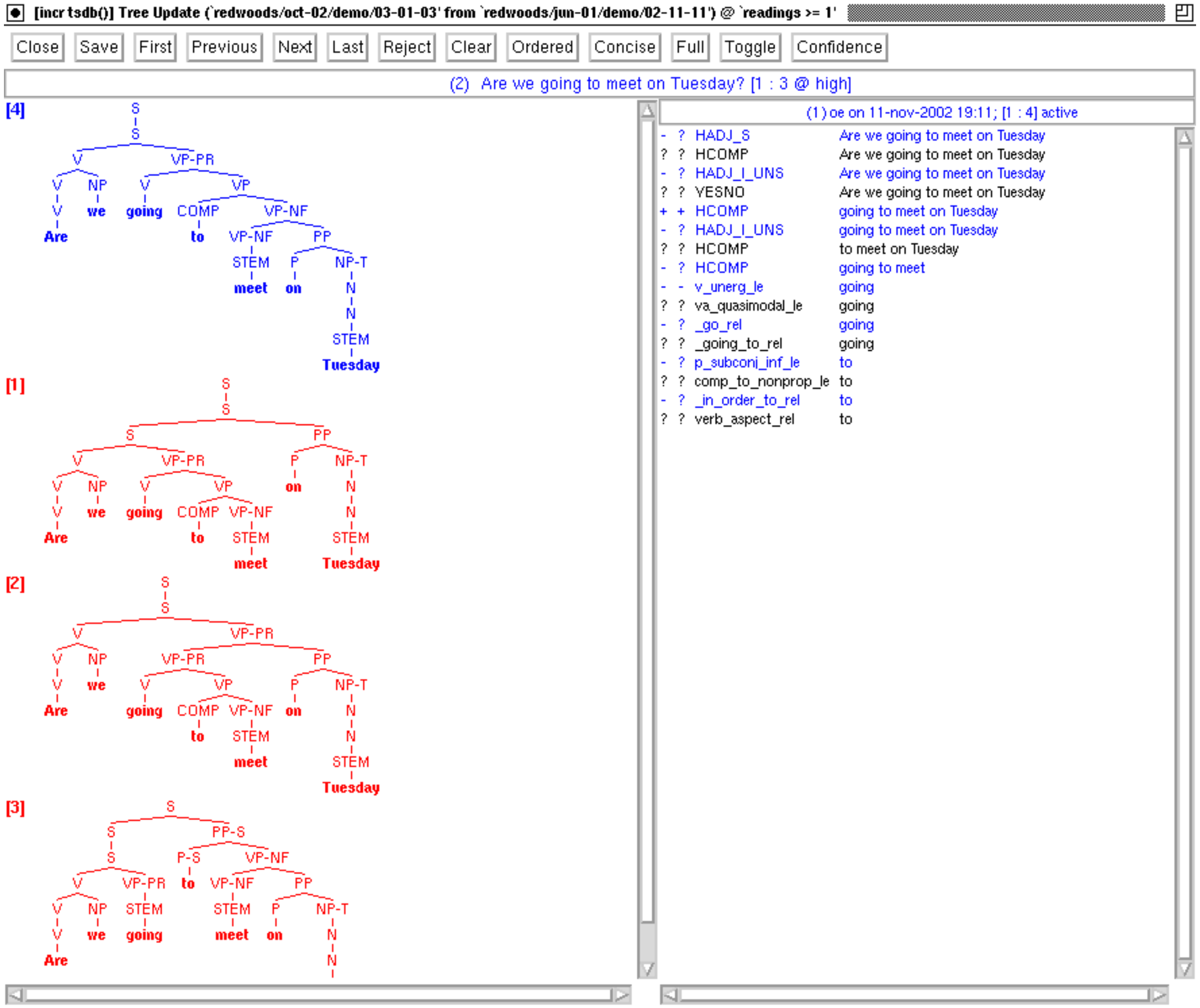
LinGO Redwoods: a Rich and Dynamic Treebank

- 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;
- 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.

LinGO Redwoods: A Quick Test Drive



[incr tsdb()] Tree Update (`redwoods/oct-02/demo/03-01-03' from `redwoods/jun-01/demo/02-11-11') @ `readings >= 1'

(2) Are we going to meet on Tuesday? [1 : 3 @ high]

(1) oe on 11-nov-2002 19:11; [1 : 4] active

```
- ? HADJ_S              Are we going to meet on Tuesday
? ? HCOMP               Are we going to meet on Tuesday
- ? HADJ_I_UNS          Are we going to meet on Tuesday
? ? YESNO               Are we going to meet on Tuesday
+ + HCOMP               going to meet on Tuesday
- ? HADJ_I_UNS          going to meet on Tuesday
? ? HCOMP               to meet on Tuesday
- ? HCOMP               going to meet
- - v_unerg_le          going
? ? va_quasimodal_le    going
- ? _go_rel             going
? ? _going_to_rel       going
- ? p_subconj_inf_le    to
? ? comp_to_nonprop_le  to
- ? _in_order_to_rel    to
? ? verb_aspect_rel     to
```

Annotation: Basic Discriminating Properties

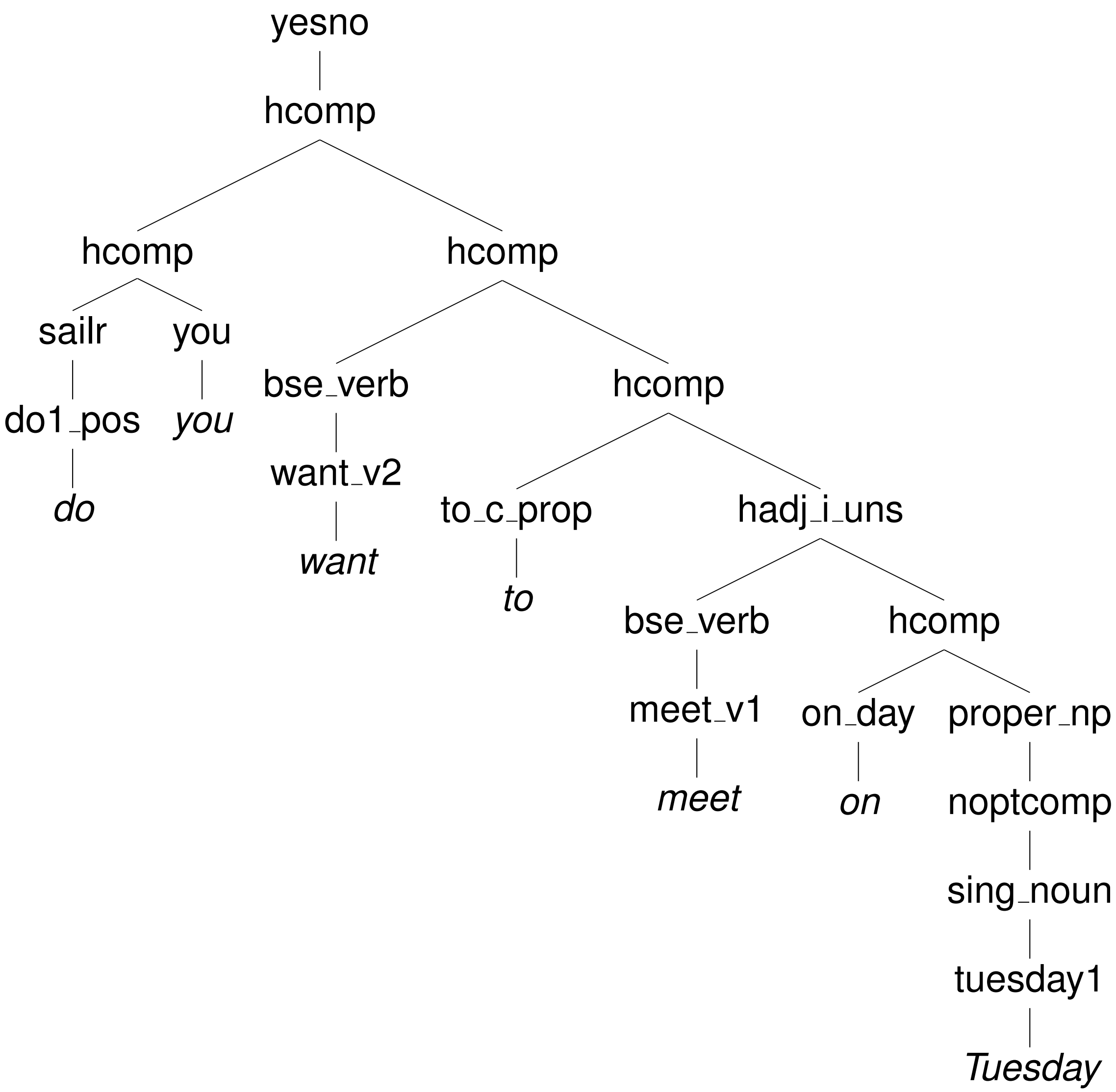
Key Notions

- 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.

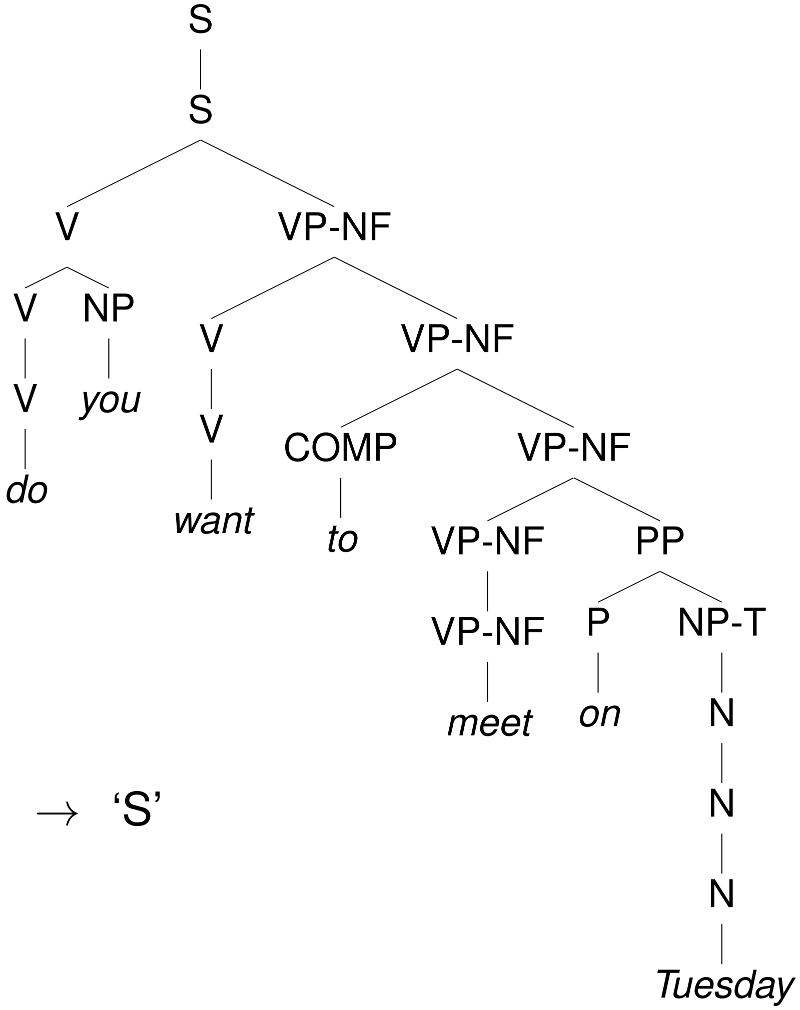
Redwoods Representations: Native Encoding



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:

$$\mathit{label}\begin{bmatrix}\texttt{SYNSEM.LOCAL.CAT}\begin{bmatrix}\texttt{HEAD}\ \mathit{verbal}\\ \texttt{VAL}\begin{bmatrix}\texttt{SUBJ} & \langle\rangle\\ \texttt{COMPS} & \mathit{*olist*}\end{bmatrix}\end{bmatrix}\end{bmatrix} \rightarrow \text{‘S’}$$



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)]
    }
```

Holistic Testing in Grammar Engineering

- Resource grammar serves multiple purposes

  Implementation of linguistic analyses

  Coverage of corpus phenomena for multiple applications

- Grammar tuning for one ‘customer’ can affect others

  Additional analyses — more ambiguity

  Dropped analyses — loss of coverage

- Exhaustive testing is required to ensure consistency

  Representative data exhibiting all relevant phenomena

  Identification of the intended analysis for each item

- Cost of testing must be kept low

  Multiple test–tune–test cycles needed for each grammar release

  Only essential discriminants presented to grammar writer

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.

Some of the Active Development Sets

	active = 0			active = 1			active > 1		
	$\sharp$	$\|$	$\times$	$\sharp$	$\|$	$\times$	$\sharp$	$\|$	$\times$
VM_6	15	14·3	8670	3811	7·9	111	0	0·0	0
$\mathrm{EC}_{\mathrm{OC}}$	38	13·1	259	1144	7·4	47	2	6·0	21
TREC	4	11·5	86	662	7·9	20	0	0·0	0
HIKE	1	22·0	876	318	12·9	187	0	0·0	0

- Variation in domain, type (spoken, email, QA, narrative), complexity;
- minor residues of rejected analyses and unresolved ambiguity;
- complemented by syntactic (1348) and semantic (107) test suites.

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

		original		matches		update			final	
Aggregate	**items**	**in**	**out**	**yes**	**no**	**in**	**out**	**new**	**in**	**out**
	$\sharp$	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 $\geq$ 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 $\geq$ 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.

Conclusions — Outlook

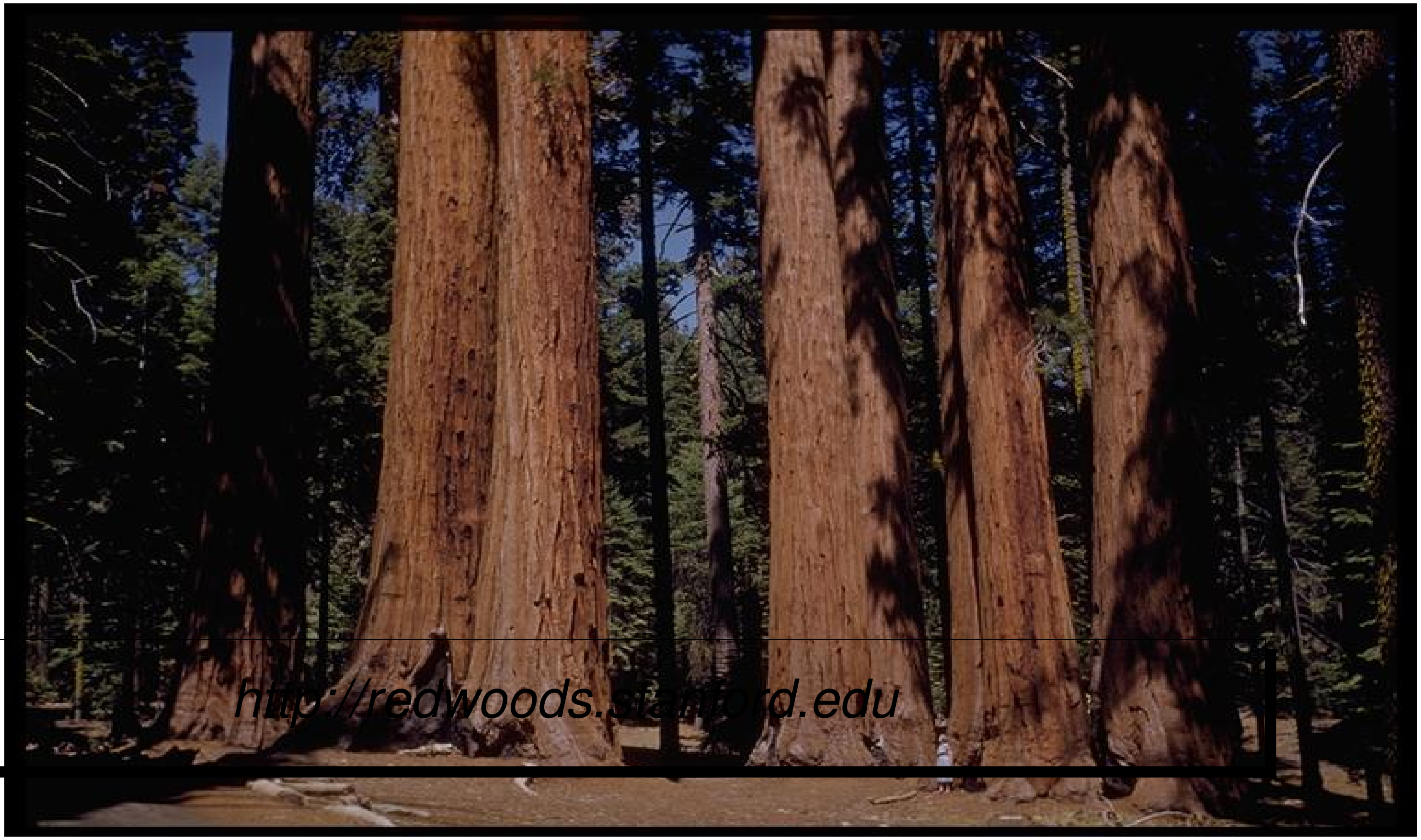
- ‘Deep’ grammar-based processing requires adequate stochastic models;
- no existing treebank resources with suitable granularity and flexibility;
- LinGO Redwoods treebank tied to broad-coverage HPSG implementation;

→ paradigm shift in sustainable, broad-coverage grammar engineering.

More Recent Developments

- Expanded annotation in multiple domains with varied characteristics;
- Japanese off-spring: *Hinoki* (NTT); 92 % coverage on dictionary definitions;
- systematic inter-annotator agreement experiments; ‘blazing’ the trail.

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.

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;

→ native encoding performs far better than labeled constituent trees.