

Topics in Computational Linguistics — Grammar Engineering —

Dan Flickinger

CSLI Stanford & Saarland University

`danf@csli.stanford.edu`

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

`oe@csli.stanford.edu`

<http://lingo.stanford.edu/courses/05/ge/>



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@csl.i.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.



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’ (i.e. intended) analysis is a very hard problem (AI);
- inclusion of (non-grammatical) sortal constraints is generally undesirable.

‘Current’ State of Affairs

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



Redwoods: Objectives and Basic Approach

Background — Motivation

- Facilitate research into stochastic disambiguation for HPSG (parsing);
- overcome limitations of existing resources (e.g. the PTB and alikes);
- define and train parse ranking model over broad-coverage grammar.

A Rich and Dynamic Treebank

- Grounded in existing broad-coverage grammar and parsing system;
- hand-disambiguate: record elementary discriminating decisions;
- provide syntactic and semantic information in variable formats;
- treebank evolution and (semi-)automatic updates with the grammar.



Deep Linguistic Processing with HPSG: DELPH-IN

Set-Up

- Loosely organized group of institutions and interested individuals;
- rooted in ‘linguistic’ NLP but geared towards practical applications;
- DELPH-IN resources in wide use: research, education, applications.

History

- Originally CSLI Stanford – Saarland University (VerbMobil, 1994);
- Tokyo University (Tsuji Laboratory): logic compilation techniques;
- Cambridge and Sussex Universities: efficient and accurate NLP;
- Trondheim and Oslo Universities, NTT: additional languages, MT.



An Open-Source Repository: Existing Resources

Common Reference Formalism

- Strongly typed, conjunctive, closed world typed feature structure logic;
- blend of [Carpenter, 1992], [Copestake, 1992], and [Krieger, 1995].

Engineering and Processing Environments

- LKB: grammar development environment (Lisp) [Copestake, 2002];
- PET: efficient, industry quality runtime engine (C⁺⁺) [Callmeier, 2000];
- [incr tsdb()]: competence and performance profiler [Oepen, 2000].

Common Grammars on Multiple Platforms

- English (CSLI): English Resource Grammar; Dan Flickinger et al.
- Japanese (DFKI, YY Technologies): Melanie Siegel & Emily M. Bender.



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.



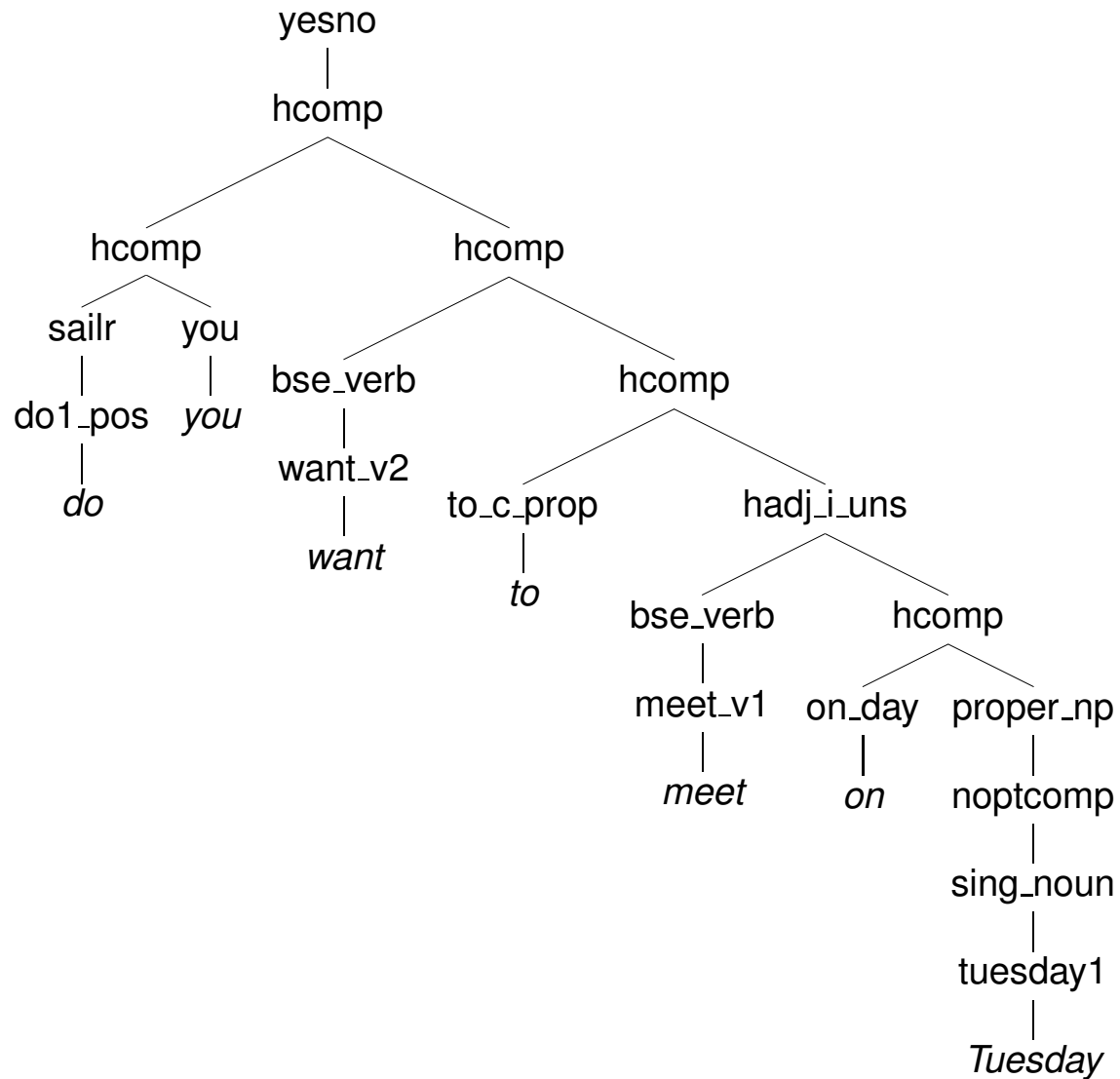
Annotation: Basic Discriminating Properties

- 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;

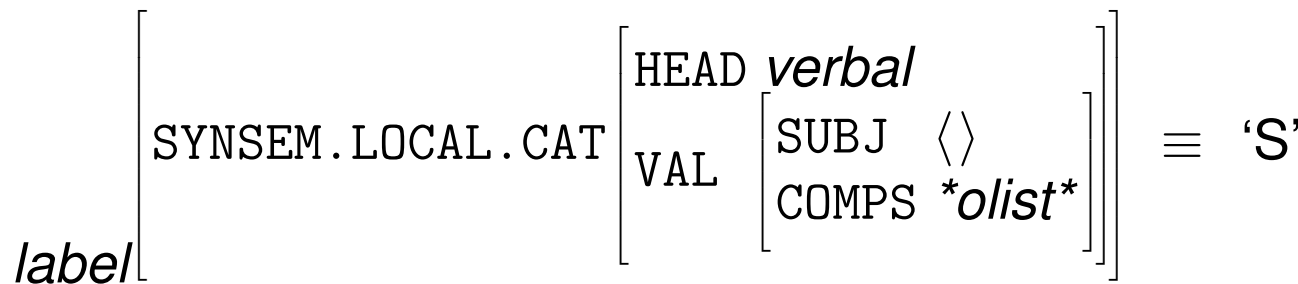
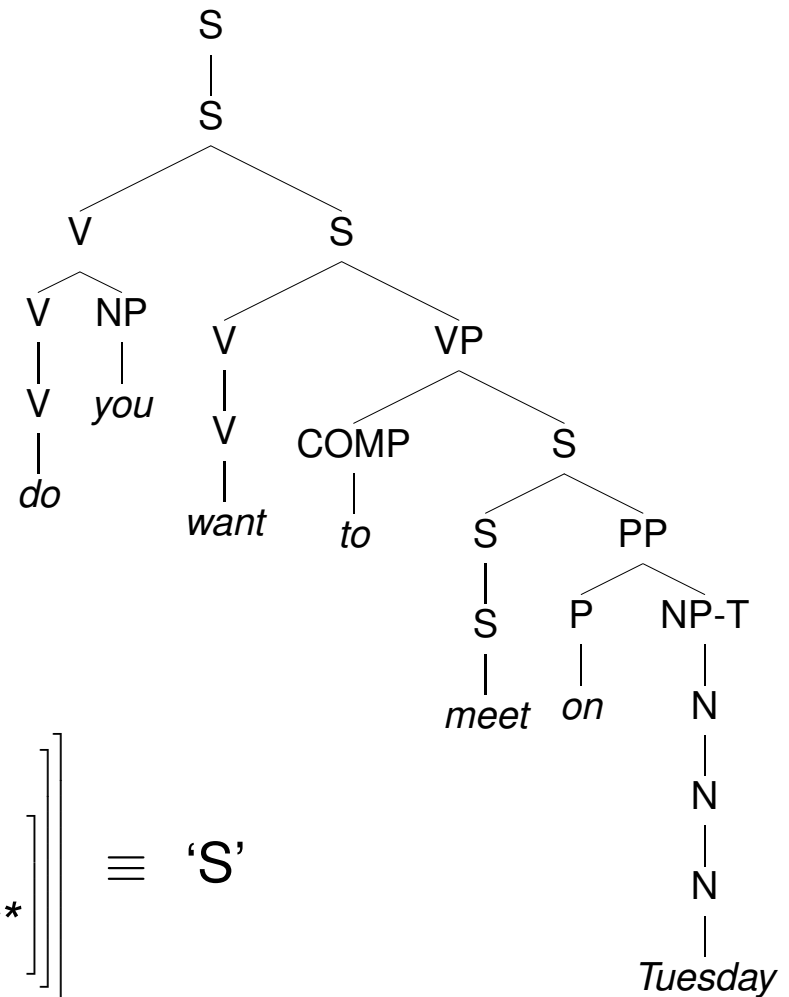


Redwoods Representations: Native Encoding



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.



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.

```
_4:{  
  _4:int_rel[SOA e2:_want2_rel]  
  e2:_want2_rel[ARG1 x4:pron_rel, ARG4 _2:hypo_rel]  
  _1:def_rel[BV x4:pron_rel]  
  _2:hypo_rel[SOA e18:_meet_v_rel]  
  e18:_meet_v_rel[ARG1 x4:pron_rel]  
  e19:_on_temp_rel[ARG e18:_meet_v_rel, ARG3 x21:dofw_rel]  
  x21:dofw_rel[NAMED :tue]  
  _3:def_np_rel[BV x21:dofw_rel]  
}
```



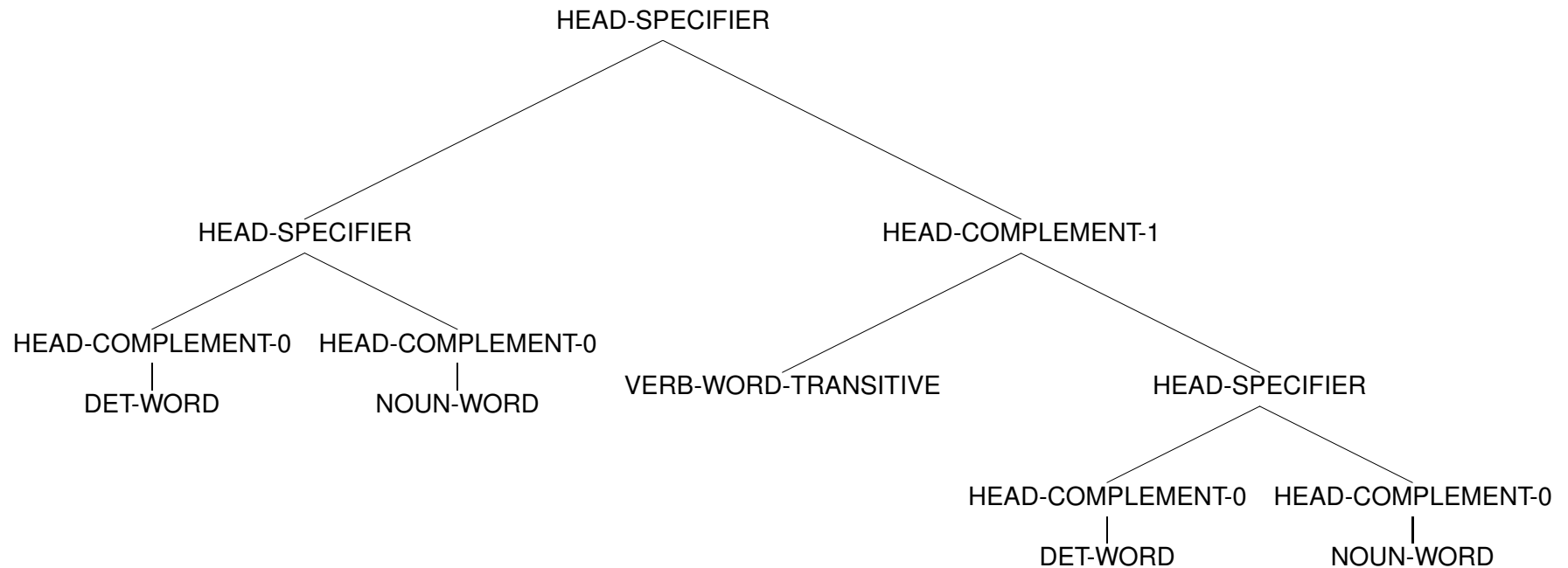
Redwoods Development Status: 3rd Growth

	all parses			active = 0			active = 1			active > 1		
	#		×	#		×	#		×	#		×
VM₆	2706	7.7	46.7	216	9.4	63.5	2482	8.3	43.5	6	15.8	757.8
VM₁₃	2279	8.5	61.9	248	10.8	80.5	2029	8.7	59.5	2	15.5	198.0
VM₃₁	1967	6.2	27.9	216	10.1	95.9	1746	7.5	30.8	5	8.4	20.8
VM₃₂	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.



(Simplified) PCFG Estimation Example



probability	context-free rule (i.e. local tree of depth one)
$2/3 = 0.66$	HEAD-SPECIFIER \rightarrow HEAD-COMPLEMENT-0 HEAD-COMPLEMENT-0
$1/3 = 0.33$	HEAD-SPECIFIER \rightarrow HEAD-SPECIFIER HEAD-COMPLEMENT-1
$1/1 = 1.00$	HEAD-COMPLEMENT-1 \rightarrow VERB-WORD-TRANSITIVE HEAD-SPECIFIER
$2/4 = 0.50$	HEAD-COMPLEMENT-0 \rightarrow NOUN-WORD
$2/4 = 0.50$	HEAD-COMPLEMENT-0 \rightarrow DET-WORD



Redwoods Applications: Parse Disambiguation

- Manning & Toutanova (Stanford): generative and conditional models;
- Baldrige & 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;
- alternatively, full parse forest (polynomial) plus selective unpacking.



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;
- *conditional* model: given some context, sample properties of events.

Conditional Parse Selection

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

$$p(t_i|s) = \frac{\exp \sum_j \lambda_j f_j(t_i)}{\sum_{k=1 \dots n} \exp \sum_j \lambda_j f_j(t_k)} \quad (1)$$



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 NeGra format;
- [Bouma et al., 2001] HPSG for Dutch, dependency structures only;
- emerging ‘HPSG’ treebanks for Bulgarian, Polish, maybe others;
- to our best knowledge, no existing *rich* and *dynamic* approach.



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/>



Outlook: Redwoods for Everyone



Towards a common, open-source basis for stochastic HPSG



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.