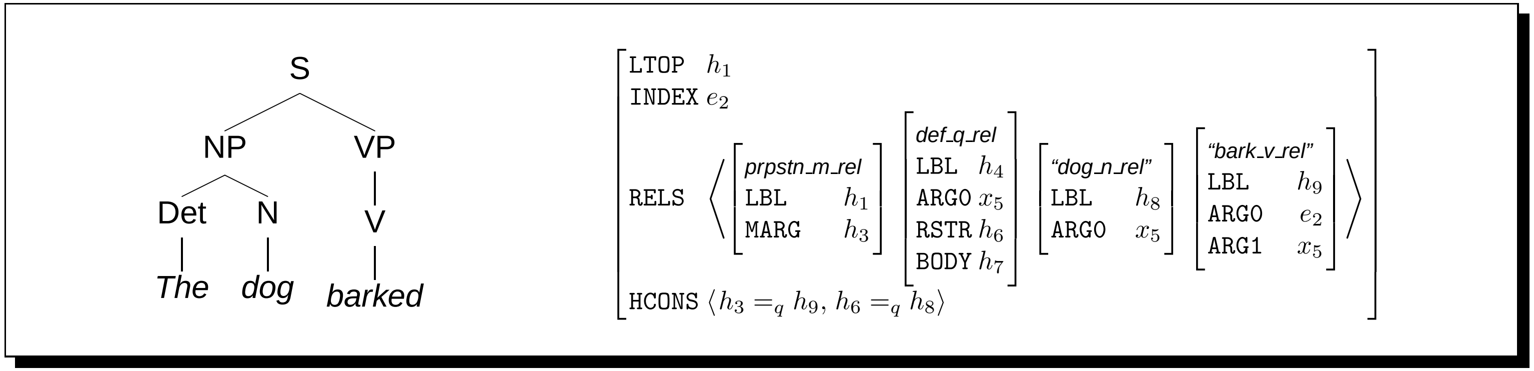


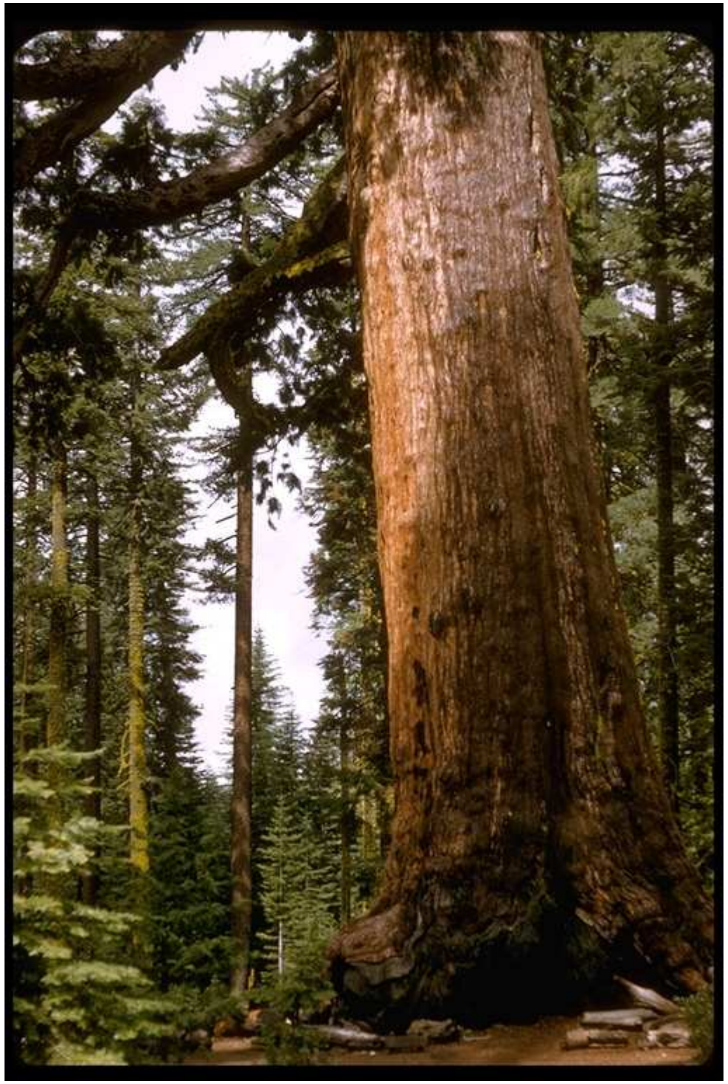
Computational Linguistics (INF2820 — Complexity)

$$\alpha \in C,\ \beta_i \in (C \cup \Sigma)^*,\ \gamma \in (C \cup \Sigma)^+,\ \delta \in \Sigma^+$$

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

`oe@ifi.uio.no`



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`

LinGO English Resource Grammar

Linguistic Grammars On-Line (`http://lingo.stanford.edu/`)

- LinGO English Resource Grammar (Dan Flickinger et al., since 1993);
- general-purpose HPSG; domain-specific lexica (some 32,000 lexemes);
- development using LKB; high-efficiency C[++] parser for applications;
- domain-specific vocabulary addition and tuning → ~85^{+}% coverage;
- average parse times: a few seconds per sentence, for Wikipedia text;

→ exact same resource used simultaneously in many (research) projects.

An Open-Source Repository (`http://www.delph-in.net/`)

- Harmonize theory, formalism, and tools: exchange ling- and software;
- world-wide initiative, now twelve languages under active development.

LinGO Redwoods: a Rich and Dynamic Treebank

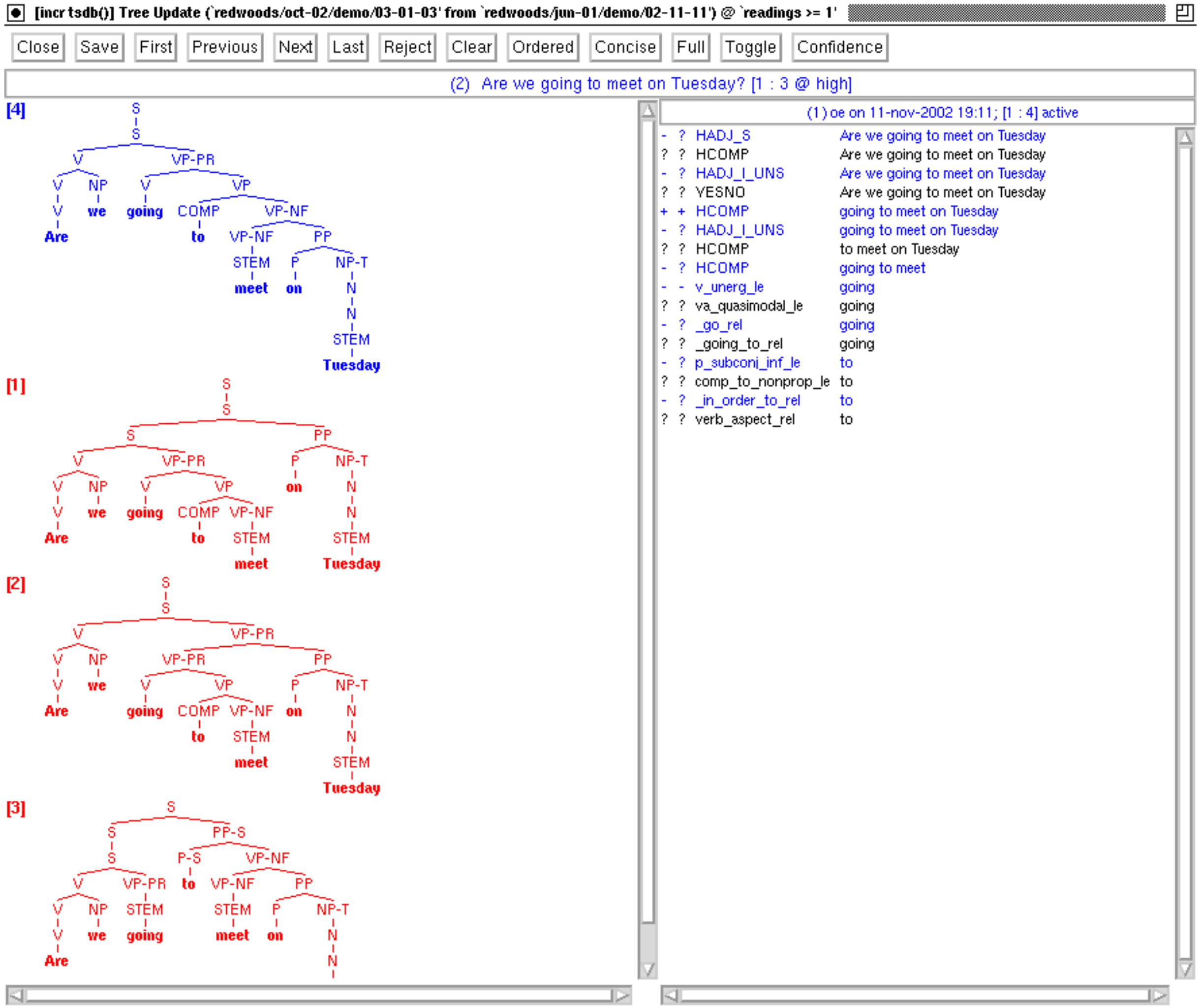
Motivation

- Broad coverage means hundreds or thousands possible analyses;
- probabilistic disambiguation for HPSG requires training material.

General Idea

- Tie treebank development to existing broad-coverage grammar;
- hand-select (or reject) intended analyses from parsed corpora;
- [Carter, 1997]: annotation using *basic discriminating* properties;
- record *annotator decisions* (and entailments) as first-class data;
- provide toolkits for dynamic mappings into various export formats.

LinGO Redwoods: A Quick Test Drive



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 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.

Review: Context-Free Grammars

- Formally, a *context-free grammar* (CFG) is a quadruple: $\langle C, \Sigma, P, S \rangle$
- C is the set of categories (aka *non-terminals*), e.g. $\{$S$,$ NP$,$ VP$,$ V$\}$;
- Σ is the vocabulary (aka *terminals*), e.g. $\{$Kim$,$ snow$,$ saw$,$ in$\}$;
- P is a set of category rewrite rules (aka *productions*), e.g.

S $\rightarrow$ NP VP
VP $\rightarrow$ V NP
NP $\rightarrow$ Kim
NP $\rightarrow$ snow
V $\rightarrow$ saw

- $S \in C$ is the *start symbol*, a filter on complete ('sentential') results;
- for each rule '$\alpha \rightarrow \beta_1, \beta_2, ..., \beta_n$' $\in P$: $\alpha \in C$ and $\beta_i \in C \cup \Sigma$; $1 \leq i \leq n$.

The Chomsky Hierarchy of (Formal) Languages

- (Formal) Languages vary in 'degree of structural complexity' exhibited;
- traditionally: a^* (iteration) vs. $a^n b^n$ (nesting) vs. $a^n b^m c^n d^m$ ('cross-serial');
- Chomsky Hierarchy: inclusion classes of formal languages; Type 0 – 3.

0	unrestricted	$\beta_1 \rightarrow \beta_2$	Turing Machine
1	context-sensitive	$\beta_1 \alpha \beta_2 \rightarrow \beta_1 \gamma \beta_2$	linearly-bounded automaton
2	context-free	$\alpha \rightarrow \beta$	push-down automaton
3	regular	$\alpha \rightarrow \delta \mid \alpha\delta$	finite-state automaton

$$\alpha \in C, \ \beta_i \in (C \cup \Sigma)^*, \ \gamma \in (C \cup \Sigma)^+, \ \delta \in \Sigma^+$$

What is the Formal Complexity of Natural Languages?

- Minimally context-free (center self-embedding, e.g. in relative clauses);
- (Culy; Shieber, 1985): *not* context-free (Bambara, Swiss German);
- (Joshi, 1985): extra class of *mildly* context-sensitive languages (TAG).