Ling/CSE 472:
Introduction to Computational Linguistics

May 14: Grammar-Based Treebanking

Overview

- Announcements: Assignment 5; Presentation schedule
- Review: Edges, nodes, consituents
- Grammar-based treebanking
    - History & Motivation
    - Contents & Methodology
    - Outlook
- Reading questions

Review: Nodes, edges, constituents

- What's the difference?
- Node: A part of a parse tree
- Edge: An object manipulated by a parser in the course of finding parse trees
  - Active v. passive edges
- Constituent: A substring of a sentence dominated by a node in a parse tree

=> Demo

HPSG in one slide

- Key references: Pollard & Sag 1987, Pollard & Sag 1994, Sag, Wasow & Bender 2003 (textbook)

- Phrase structure grammar: Like CFG but with elaborate feature structures instead of atomic node labels

- Monostratal/surface oriented: One structure per input item (no movement), with both syntactic and semantic information

- Lexicalist: Rich information in lexical entries (+ type hierarchy to capture generalizations)

- Core & periphery: Construction inventory includes both very general and very idiosyncratic rules

Flickinger et al 2017: Central claims

- Developing complex linguistic annotations calls for an approach which allows for the incremental improvement of existing annotations by encoding all manual effort in such a way that its value is preserved and enhanced even as the resource is improved over time

- Manual effort:
  - Annotation design => Encode in a grammar
  - Disambiguation => Store disambiguation decisions in a treebank

Minimal Recursion Semantics in one slide

- Key references: Copestake et al 2005, Bender et al 2015
- Underspecified description of logical forms
- Captures predicate-argument structure, partial constraints on quantifier scope, morpho-semantic features
- Computationally tractable, grammar-compatible, and linguistically expressive

English Resource Grammar (Flickinger 2000, 2011)

- Under continuous development since 1993.
- 38,000 item lexicon: function words, open-class words with ‘non-standard’ properties
- 1214 release: 225 syntactic rules, 70 lexical rules, 975 leaf lexical types
- Unknown words given default lexical entries based on POS tagging
- 85-95% coverage of open domain, well-edited English text
- Development genres: newspaper text, Wikipedia pages, bio-medical research, literature, customer service emails, meeting scheduling dialogues…

Redwoods Treebank (Oepen et al 2004)

- Under development since 2001
- As of ‘ninth growth’, 1.5 million tokens
- Initial motivation: train parse ranking models
- Also quite useful for grammar maintenance and development

Redwoods: Contents

- Rich syntactico-semantic structures, from which different ‘views’ can be projected.

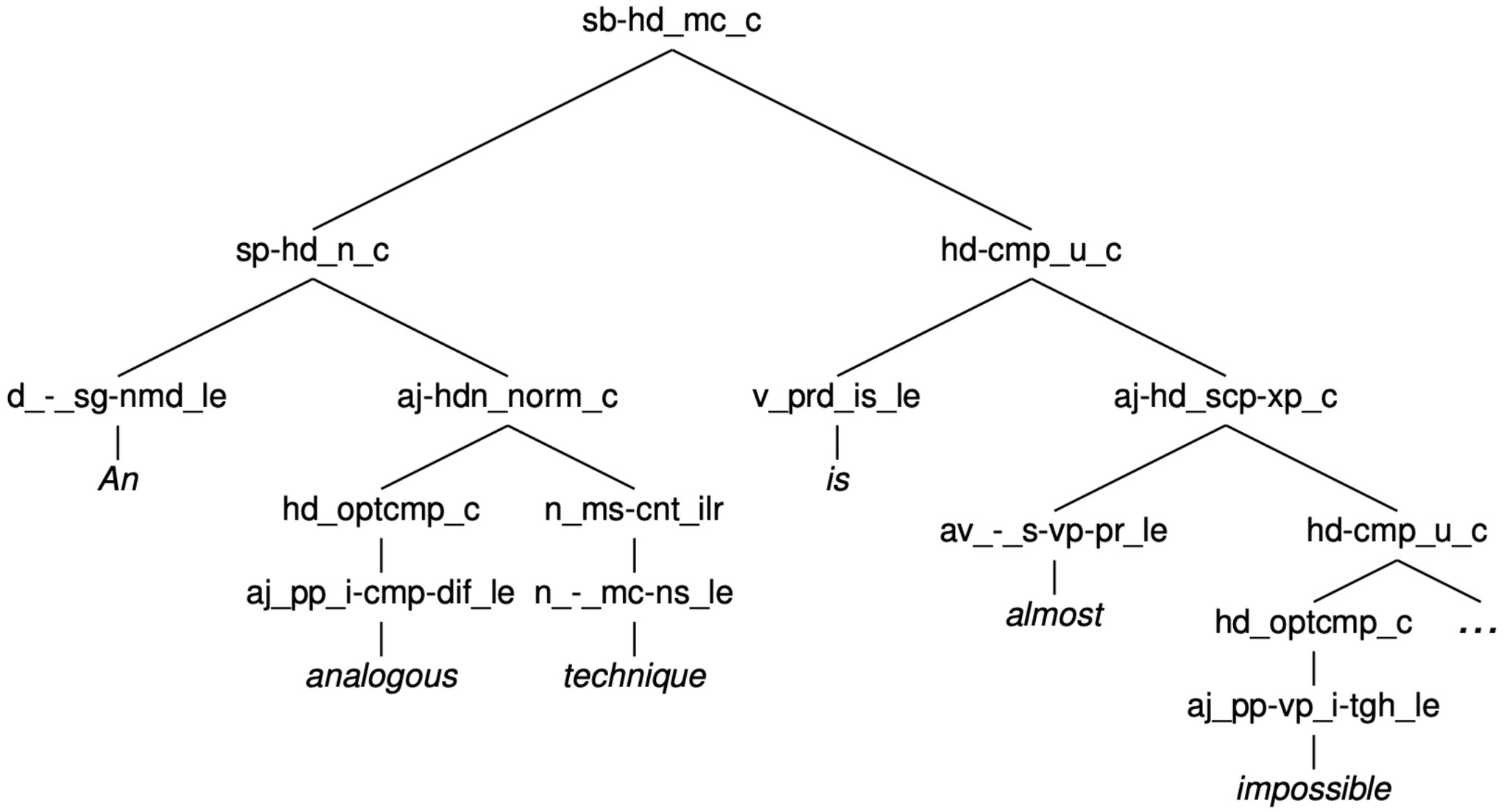
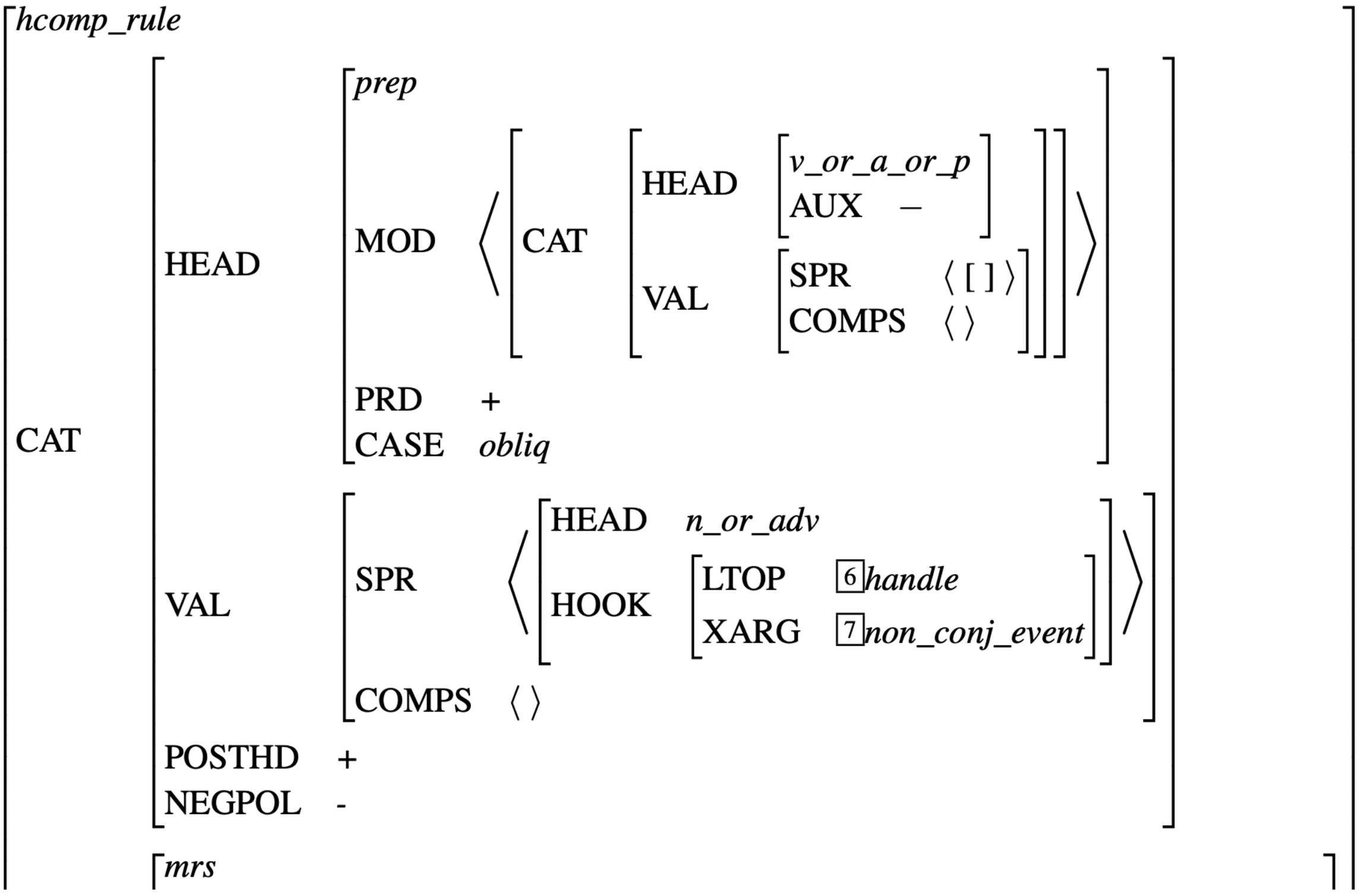
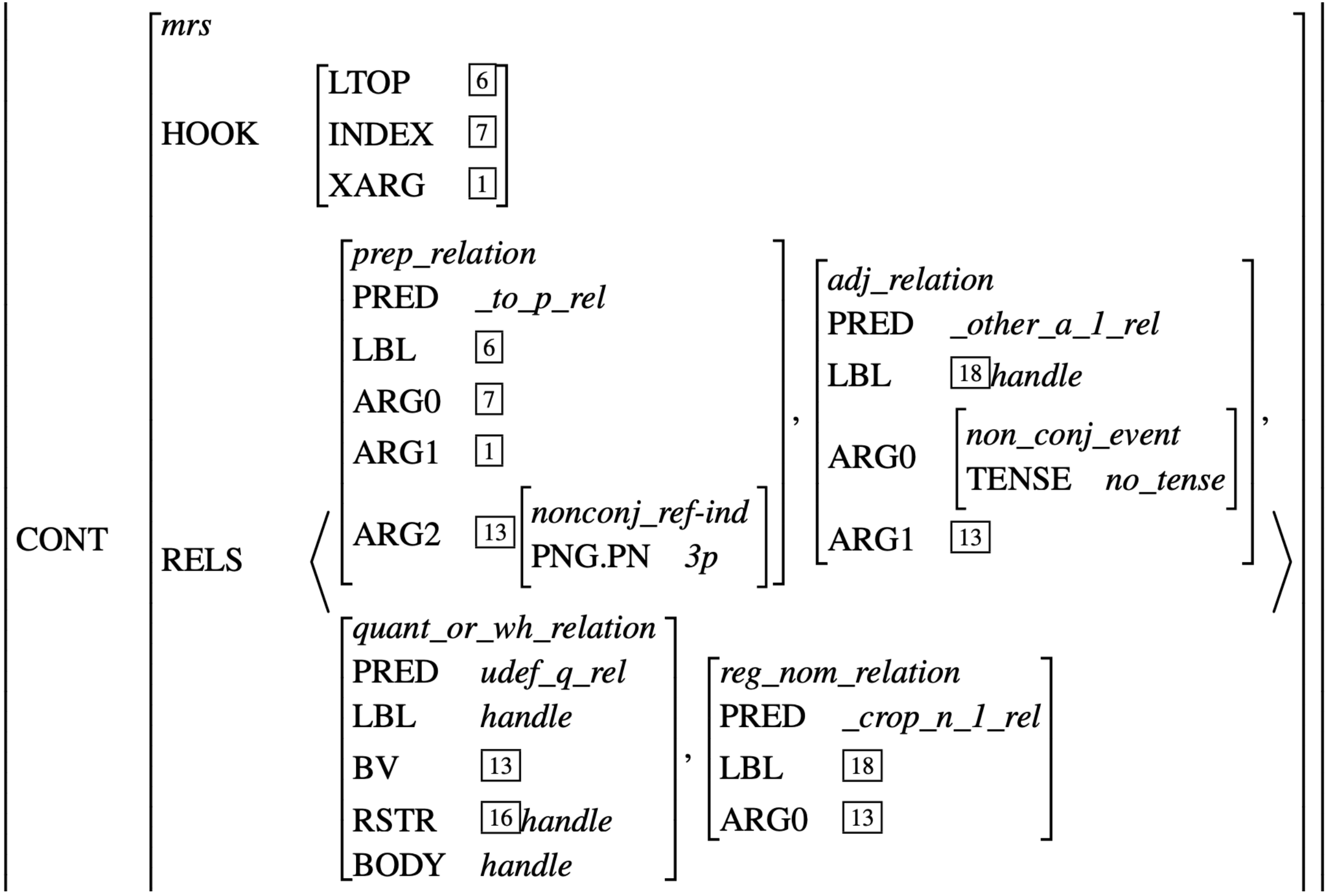


Fig. 1 ERG derivation tree for example (1).


hcomp_rule
CAT
HEAD
prep
MOD
CAT
HEAD
v_or_a_or_p
AUX –
VAL
SPR ⟨ [ ] ⟩
COMPS ⟨ ⟩
PRD +
CASE obliq
VAL
SPR
HEAD n_or_adv
HOOK
LTOP 6 handle
XARG 7 non_conj_event
COMPS ⟨ ⟩
POSTHD +
NEGPOL -
mrs


mrs
HOOK
LTOP 6
INDEX 7
XARG 1
CONT
RELS
prep_relation
PRED _to_p_rel
LBL 6
ARG0 7
ARG1 1
ARG2 13
nonconj_ref-ind
PNG.PN 3p
adj_relation
PRED _other_a_1_rel
LBL 18 handle
ARG0
non_conj_event
TENSE no_tense
ARG1 13
quant_or_wh_relation
PRED udef_q_rel
LBL handle
BV 13
RSTR 16 handle
BODY handle
reg_nom_relation
PRED _crop_n_1_rel
LBL 18
ARG0 13

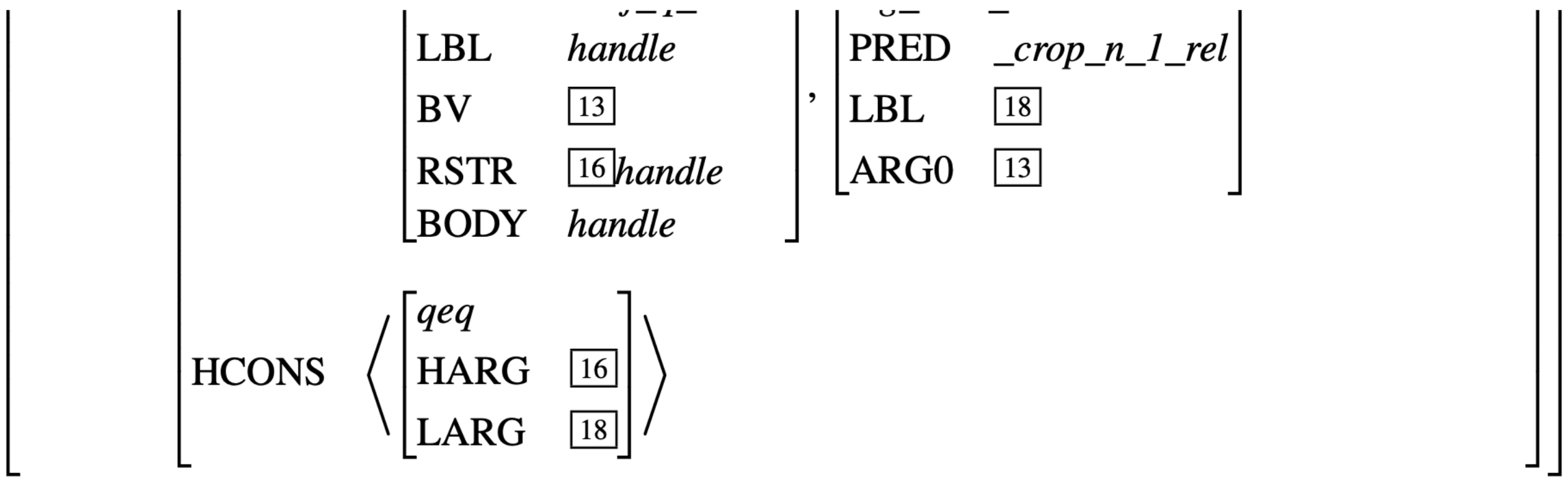


Fig. 2 Partial feature structure for PP *to other crops*

$$
\langle h_1,
\left|\begin{array}{l}
h_4{:}\_a\_q(\text{BV } x_6, \text{RSTR } h_7, \text{BODY } h_5),\\
h_8{:}\_analogous\_a\_to(\text{ARG0 } e_9, \text{ARG1 } x_6), h_8{:}comp(\text{ARG0 } e_{11}, \text{ARG1 } e_9, \text{ARG2 } \_),\\
h_8{:}\_technique\_n\_1(\text{ARG0 } x_6),\\
h_2{:}\_almost\_a\_1(\text{ARG0 } e_{12}, \text{ARG1 } h_{13}), h_{14}{:}\_impossible\_a\_for(\text{ARG0 } e_3, \text{ARG1 } h_{15}, \text{ARG2 } \_),\\
h_{17}{:}\_apply\_v\_to(\text{ARG0 } e_{18}, \text{ARG1 } \_, \text{ARG2 } x_6, \text{ARG3 } x_{20}),\\
h_{21}{:}udef\_q(\text{BV } x_{20}, \text{RSTR } h_{22}, \text{BODY } h_{23}), h_{24}{:}\_other\_a\_1(\text{ARG0 } e_{25}, \text{ARG1 } x_{20}),\\
h_{24}{:}\_crop\_n\_1(\text{ARG0 } x_{20})
\end{array}\right|\\
\{h_1 =_q h_2, h_7 =_q h_8, h_{13} =_q h_{14}, h_{15} =_q h_{17}, h_{22} =_q h_{24}\}\rangle
$$

Fig. 3 Minimal Recursion Semantics for example (1).

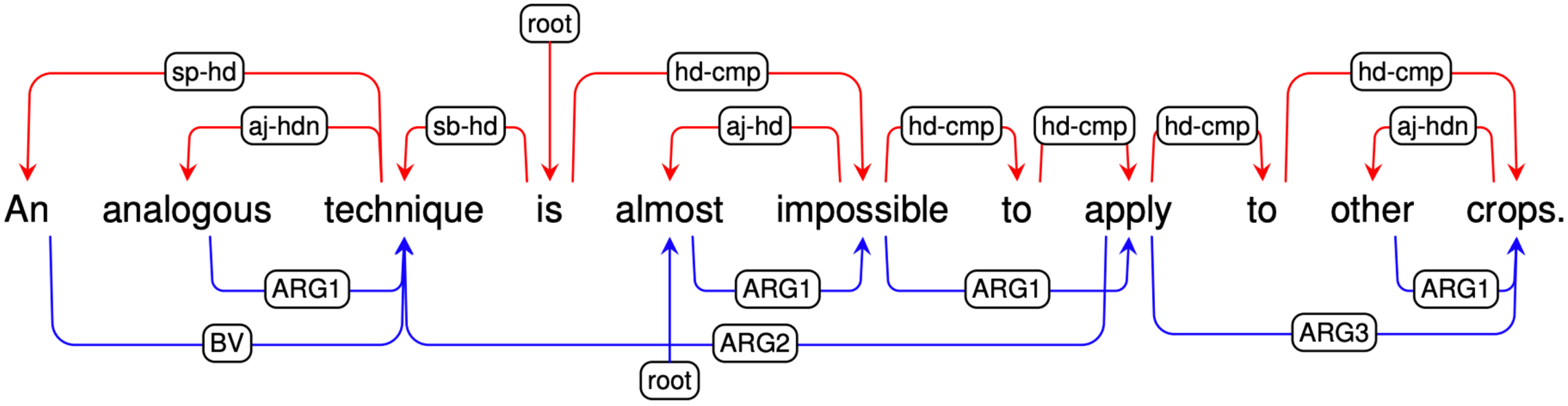


Fig. 4 Bi-lexical syntactic and semantic dependencies for (1).

Redwoods: Methodology

- Parse input corpus
- Calculate ‘discriminants’: properties shared by only a subset of the trees in parse forest (Carter 1997)
    - Picking one tree from among thousands or millions would be infeasible
    - Drawing trees with that level of detail would be infeasible
    - Picking discriminants is quite doable!
- Store both resulting tree & discriminants chosen (and inferred)
    - Maximum value out of all human annotator time

Very high inter-annotator agreement
=> very consistent annotation

- From Bender et al 2015, over 150 sentences from *The Little Prince*

	Annotator Comparison			
Metric	**A vs. B**	**A vs. C**	**B vs. C**	**Average**
Exact Match	0.73	0.65	0.70	0.70
EDM_a	0.93	0.92	0.94	0.93
EDM_{na}	0.94	0.94	0.95	0.94

Table 1: Exact match ERS and Elementary Dependency Match across three annotators.

- Comparable metric for AMR over the same data is 0.71
  "SMATCH" (comparable to EDM) (Banarescu et al 2013)

Dynamic treebanking

- Dynamic *refinement* of the treebank
    - Parse corpus with new grammar (better coverage, improved representations)
    - Rerun discriminants chosen in previous annotation rounds
    - Address remaining added ambiguity / newly parsed sentences
- Dynamic *extension* of the treebank
    - Linguistic analysis encoded as a grammar (as opposed to annotation guidelines) can be automatically deployed to new text

Redwoods: Outlook

- Switch from treebanking based on top-500 parses to full-forest (Packard 2015)

- Treebanking over robust parsing strategies to capture the remaining 5-15% of sentences

- Integrating further kinds of linguistic annotations (coreference, fine-grained word sense, information structure…)

Reading questions

- Can't probability at times solve issues for undesirable ambiguity? Like in the sentence, "The dog barks", it is more likely for the sentence to be a NP VP instead of a noun-noun compound so can't the latter be ruled out?

- Does treebanking ERG make it possible to give probability to different parse trees? Is that how it's different from the Penn Tree Bank?

Reading questions

- The reading mentions ERG lexical type that extends the traditional part of speech tag, which serves as a recipe, and combined with the grammar to regenerate the HPSG analysis. So the added complexity to the types of annotation used increases the information that can be obtained from the input, and the text mentions the size and scope of the analysis and how comprehensive it is. Does this mean the annotation scheme and the method presented in the text is a better way of representing much fuller information from a given input (and in turn, is a better way of generating a feature-structure), or it serves a different purpose compared to the ones introduced in the lectures?

- The authors say that "our purpose in providing this short tour of a feature structure has been to illuminate the level of detail involved in both the grammar and the resulting representations". Does it mean that the authors argue that the annotator should not conduct an extremely detailed grammar annotation for the treebank?

Reading questions

- Not sure I understand exactly what methodology the authors are proposing. Manual, human annotations but supported by machines? How is that done?

- In presenting a binary choice between discriminants, how does the computer decide which two discriminants to present to the annotator? Is it based on probability?

- The reading mentions how TreeBanker paid attention to annotations that could be easily judged by non-specialists. Does Redwoods do something similar?

- Redwoods was mentioned to support/been attempted for other languages. Do the methods employed translate well to code mixed corpora? Either as a "language" itself or by combining versions "tuned" for independent languages.

Reading questions

- This model appears to rely heavily on human annotators, which makes sense considering the semantics and syntactic complexity of language. However, what is done if annotators don't agree on something such as when a semantic distinction is irrelevant? How can data be correctly and consistently annotated in these cases?

- I do not fully understand cases where there are multiple candidate analyses for some linguistic construction that do not vary semantically. Does this not yield multiple semantic representations?

(2) a. They will take a cab if the plane arrives late.

b. They will take a cab if it's late and ride the bus if it's on time.

c. They will take a cab and we'll call our friends if it's late.

Reading questions

- This reading posed some challenges for the ERG including the fact that "implementing a linguistic theory will fail to provide a full correct analysis of some sentences in a corpus of any size" due to the productivity of language and errors in analyzing other sentences. What steps and efforts have been taken to improve this (whether with machine learning, manual labeling, etc.)?

- Would it be make sense for multiple parsers (which perhaps focus on different domains) using a shared output format to contribute to the development of a combined treebank? I'm thinking that the treebank could have higher coverage, but some of the mutual benefit between grammar building and treebanking like using the treebank as a source of regression tests might be lost.

Reading questions

- Is it actually possible to have a truly "entirely-annotated" corpus that reaches a final stage of completion? Are there enough gold standards in all of the fields of annotation for a corpus annotation to not constantly be changing and debated? Does that matter in the scope of the goal of this paper?

- The paper cites “fine-grained word senses, anaphoric co-reference within and across sentences, information structure, and discourse relations” as types of annotation that could be introduced into Redwoods in the future. Are these subsumed under broader semantic annotation, or do they demand a separate level of annotation? Also, are the pieces of text in Redwoods all big enough that discourse marking could be done usefully with them?

Reading questions

- The reading discusses how the amount of analyses requiring manual annotation is growing. Would this amount possibly reach a point where treebanking loses its advantage of cost-effectiveness? And it would be better to use other methods?

- In the further challenges section, the paper mentions that the grammar-centric method will fail to provide a full correct analysis for some sentence in any corpus due to linguistics limitations. Whereas a CFG trained on a large corpus will be able to give a better approximation. Are probabilistic methodologies always used in practice due to their ability to approximate things?