Treebanks and HPSG

Workshop on Treebanking, NAACL-HLT, Rochester, 26 April 2007

Dan Flickinger

CSLI Stanford
danf@csli.stanford.edu
Treebanks for HPSG: Usually grammar-derived

- Derived from parsed corpora using hand-built-grammars:
  - English: Redwoods (25,000 items in three domains)
  - Japanese: Hinoki (100,000 items in one domain)
  - German: (15,000 items in one domain)
- Discriminant-based annotation tool to identify correct parse
- References, data, and software:
  www.delph-in.net
Desirable information for grammar extraction

- Fine-grained lexical subcategorization
  - 300 verb types
  - 200 noun types
  - 100 adjective types
- Complement/adjunct distinction
- Specifier/adjunct distinction
  - *the ten-foot tall tree*
  - *a much larger room*
- Compositional semantics (at each node)
Challenges for consistent annotation

- PP attachment: underspecification or ambiguity packing?
  
  *We reserved a room for Browne.*
  
  *The best route is from on top of the hill near the cave across the valley.*

- Noun-noun compounds: ambiguity again
  
  *airline reservation counter*

- Empty vs contentful prepositions
  
  *that picture of me/mine*
  
  *their rejection of the proposal*
  
  *It turned out that they won*

- Gerunds: nominal or verbal
  
  *Their singing was surprising*
Challenges for consistent annotation (cont.)

- Attributive participles
  - the sleeping dog/pill
  - the closed/open/opened window

- Adjectives or nouns?
  - the red car
  - the east wing

- Complex modifiers
  - the three-legged chair
  - the snow-covered mountain

- Discourse connectives
  - So we left
  - Well, we have two choices.
  - Then let’s meet at five.
Desirable treebank annotations

- Marked word order
  - Heavy NP-shift: *We saw on Tuesday a most amazing film.*
  - Relative clause extraposition: *Someone walked in who I hadn’t met.*
  - Locative inversion: *In the corner stood an antique coatrack.*

- Expletive elements
  - *It took ten minutes to park the car.*
  - *In the corner there is an old desk.*

- Idioms

- Words with spaces: *(in spite of)*

- Flagging errors in text (especially for non-native text)
  - *in the the room*
Potential benefits of deriving treebank from grammar

- More consistency in annotation
  Every word and phrase gets a legal annotation
- Richer data for statistical parse selection
  Alternative grammatical analyses included
- More precision in extracted grammars
  Treebank includes all and only well-formed analyses
- Compositional (flat) semantics
  Every node in a tree has semantic annotation
- Maintainability: semi-automatic re-annotation
  Less costly improvements/corrections in annotations
Potential disadvantages of grammar-derived treebanks

- Robustness
  Currently missing good analyses for 5–20% of a given corpus
- Idiosyncratic analyses for some phenomena
  *I think that they should leave.*
  *They weren’t here that week.*
- Lack of lexical semantics, anaphor binding, pragmatics, ...