Combining Structural and Statistical Information
Relevance for Efficient Processing

Dagstuhl Seminar on Efficient Language Processing
with High-level Grammar Formalisms

Aravind Joshi and Anoop Sarkar
Dept. of Computer and Information Science
University of Pennsylvania

{joshi,anoop}@linc.cis.upenn.edu

Working with a Wide-Coverage Lexicalized Grammar
The XTAG Project

Dagstuhl Seminar on Efficient Language Processing
with High-level Grammar Formalisms

Aravind Joshi and Anoop Sarkar
Dept. of Computer and Information Science
University of Pennsylvania

{joshi,anoop}@linc.cis.upenn.edu

Results of Corpus Based Error Analysis

Rank	No of errors	Category of error
#1	11	Parentheticals and appositives
#2	8	Time NP
#3	8	Missing subcat
#4	7	Multi-word construction
#5	6	Ellipsis
#6	6	Not sentences
#7	3	Relative clause with no gap
#8	2	Funny coordination
#9	2	VP coordination
#10	2	Inverted predication
#11	2	Who knows
#12	1	Missing entry
#13	1	Comparative?
#14	1	Bare infinitive

Additions to XTAG as a result of Corpus Based Error Analysis

- Parentheticals and appositives along with a treatment of punctuation.
- Time NPs
- Multi-word prepositions
- Gapless relative clauses
- Bare infinitives
- Missing subcategorization for some predicates
- Missing lexical entries
- Comparatives without ellipsis

Phenomena Covered by the XTAG English Grammar

adjuncts	infinitives
appositives	inversion
auxiliaries	it-clefts
auxiliary contractions	multi-word prep & adv
bare infinitives	negation
clausal adjuncts	noun-noun modification
copular constructions	particle movement
determiner sequencing	passives
ECM	punctuation
ergatives	quoted speech
genitives	raising
gerunds	relative clauses
imperatives	small clauses
infinitives	time NPs
inversion	topicalization
it-clefts	wh- questions
PRO control	comparatives
resultatives	idioms

http://www.cis.upenn.edu/~xtag/ has the various XTAG tools and a full online description of the XTAG English Grammar