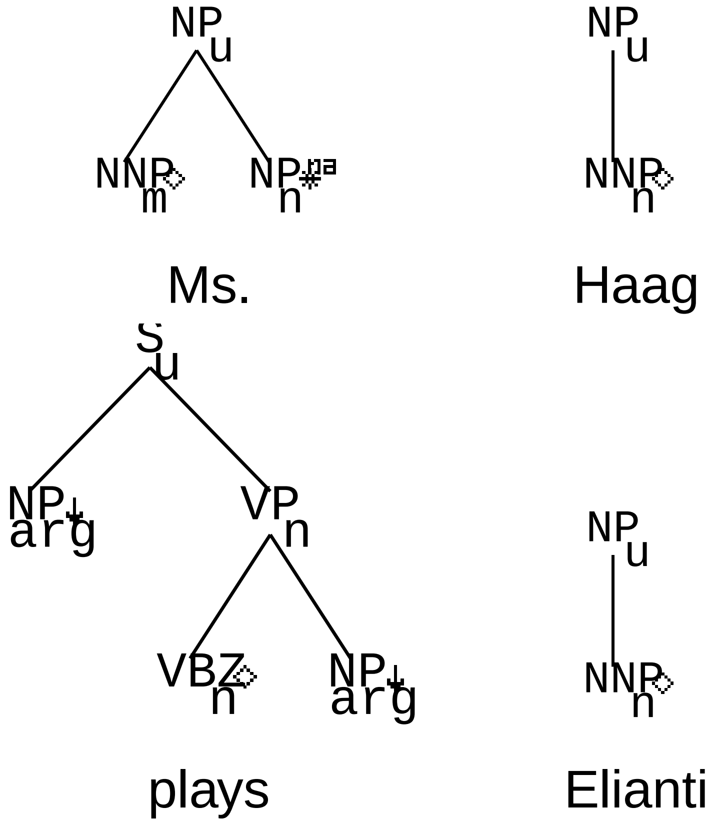
Some Experiments on Indicators of Parsing Complexity for Lexicalized Grammars

Anoop Sarkar, Fei Xia and Aravind Joshi
Dept. of Computer and Information Sciences
University of Pennsylvania
{anoop,fxia,joshi}@linc.cis.upenn.edu

Lexicalized Tree Adjoining Grammars



These trees can be combined to parse the sentence *Ms. Haag plays Elianti*.

Important Properties of LTAG wrt Parsing

- Predicate-argument structure is represented in each elementary tree.

- Adjunction instead of feature unification.

- No recursive feature structures. FSs are bounded.

Important Properties of LTAG wrt Parsing

- Transformational relations for the same predicate-argument structure are precomputed.

- Each predicate selects a family of elementary trees.

- Different sources of issues for parsing efficiency.

Parsing Efficiency

- Parsing accuracy: Evaluations done in previous work.
- Parsing efficiency: observed time complexity for producing all parses.
- The usual notion: compare different parsing algorithms wrt time, space, number of edges, . . .
- This paper: explore parsing efficiency from a viewpoint that is independent of a particular parsing algorithm.

Parsing Efficiency

- Not an alternative to comparision of parsing algorithms.

- An exploration of parsing efficiency from the perspective of a fully lexicalized grammar.

- Sources of parsing complexity that are part of the input to the parsing algorithm.

Parsing Efficiency

- We explore two issues: syntactic lexical ambiguity and clausal complexity.

- The contention: for LTAGs these issues are relevant across *all* parsing algorithms.

Experiment: The Parser

- Implementation of head-corner chart-based parser.

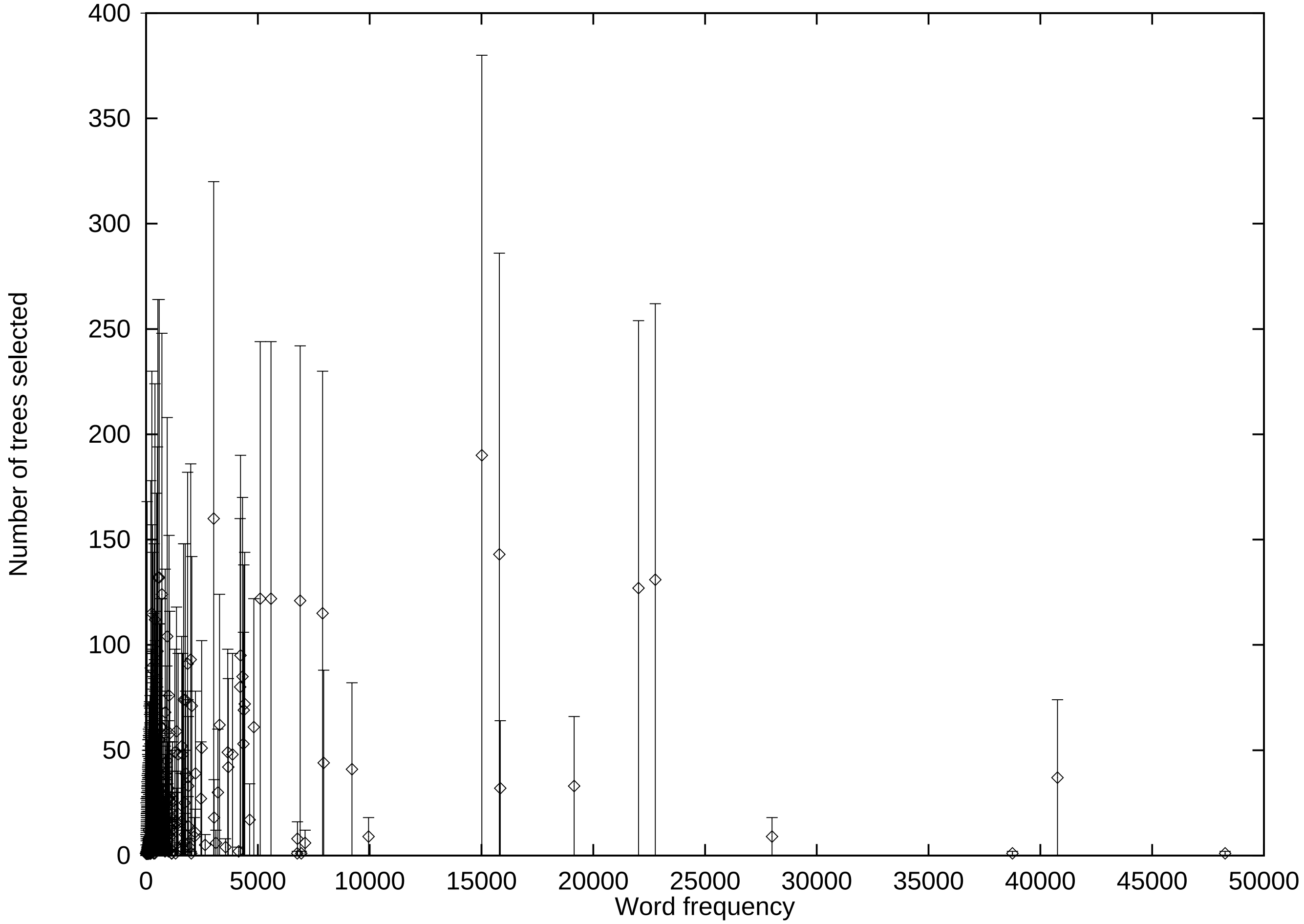
- It is bi-directional – van Noord style.

- Produces a derivation forest as output.

- Written in ANSI C: α-version available at `ftp://ftp.cis.upenn.edu/xtag/pub/lem`

Experiment: Input Grammar

- Treebank Grammar
- extracted from Sections 02–21 WSJ Penn Treebank
- 6789 tree templates, 123039 lexicalized trees
- number of word types in the lexicon is 44215
- average number of trees per word is 2.78



Number of trees selected by the words grouped by word frequency

Treebank Grammar and XTAG English Grammar

- Compared TG with the XTAG Grammar which has 1004 tree templates, 53 tree families and 1.8 million lexicalized trees.

- 82.1% of template tokens in the Treebank grammar match a corresponding template in the XTAG grammar

- 14.0% are covered by the XTAG grammar but the templates look different because of different linguistic analyses

Treebank Grammar and XTAG English Grammar

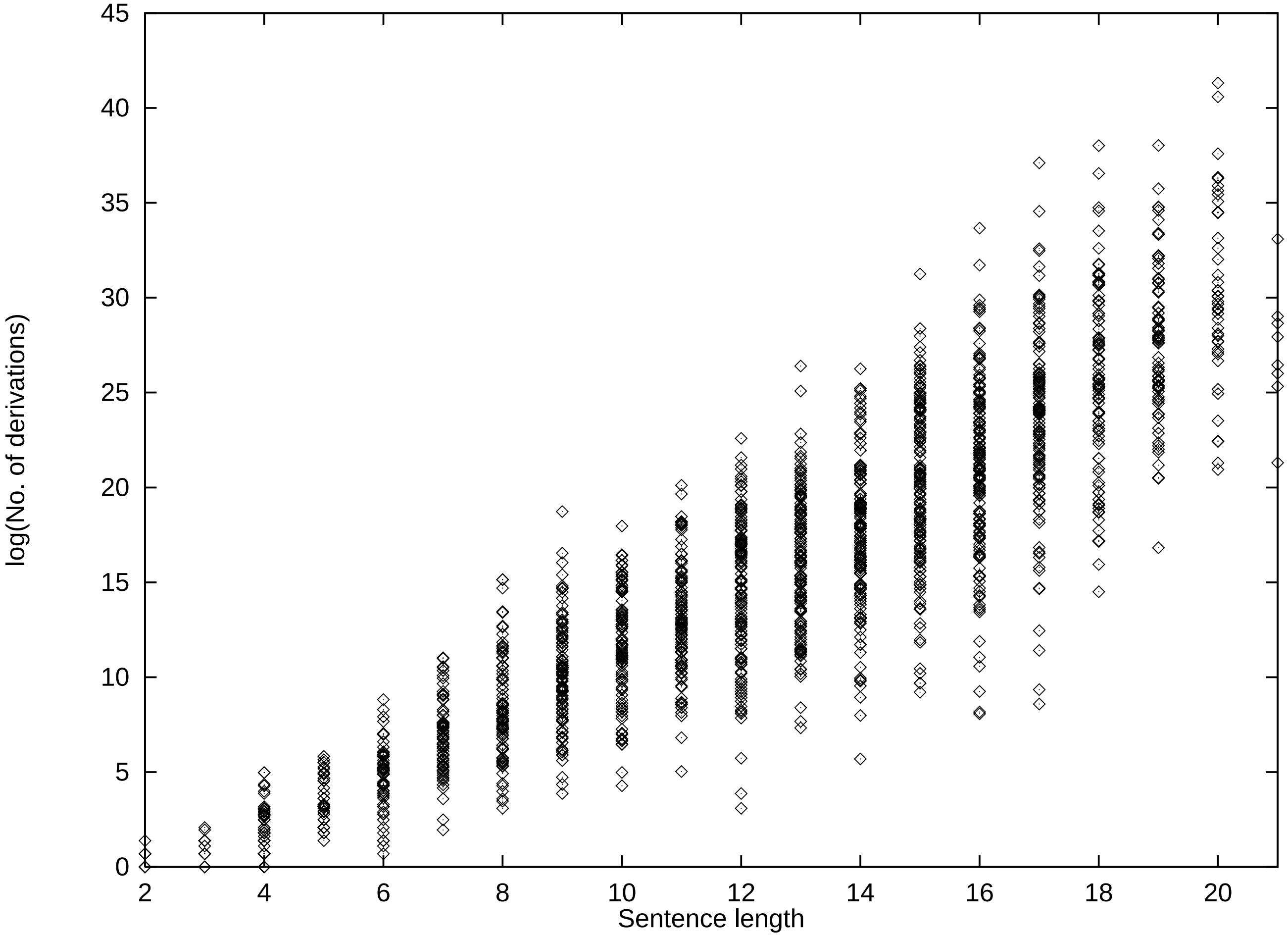
- 1.1% of template tokens in the Treebank grammar are due to annotation errors

- The remaining 2.8% are not currently covered by the XTAG grammar

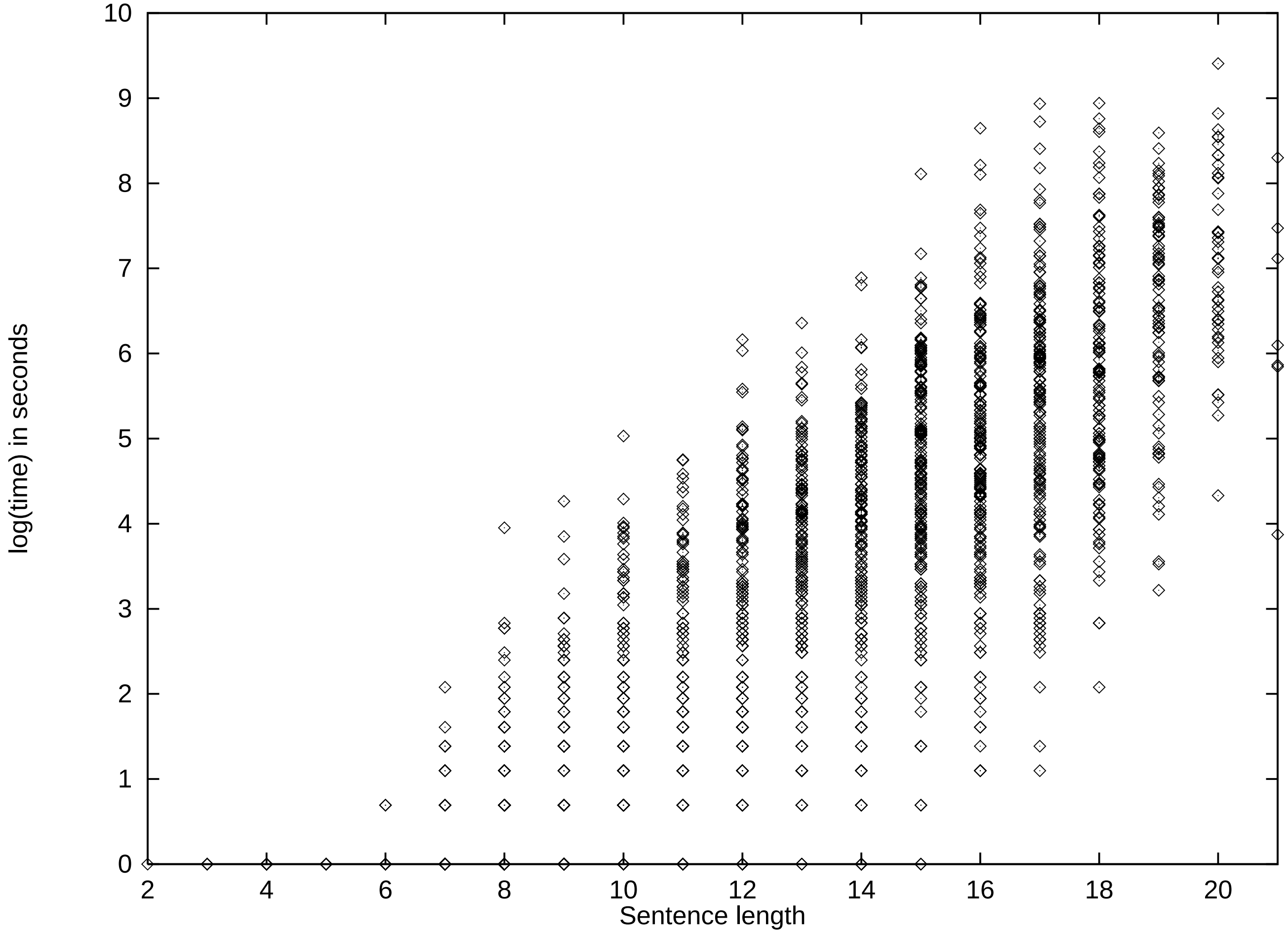
- A total of 96.1% of the structures in the Treebank grammar match up with structures in the XTAG grammar.

Experiment: Test Corpus

- input was a set of 2250 sentences
- each sentence was 21 words or less
- avg. sentence length was 12.3
- number of tokens = 27715
- output: shared forest of parses

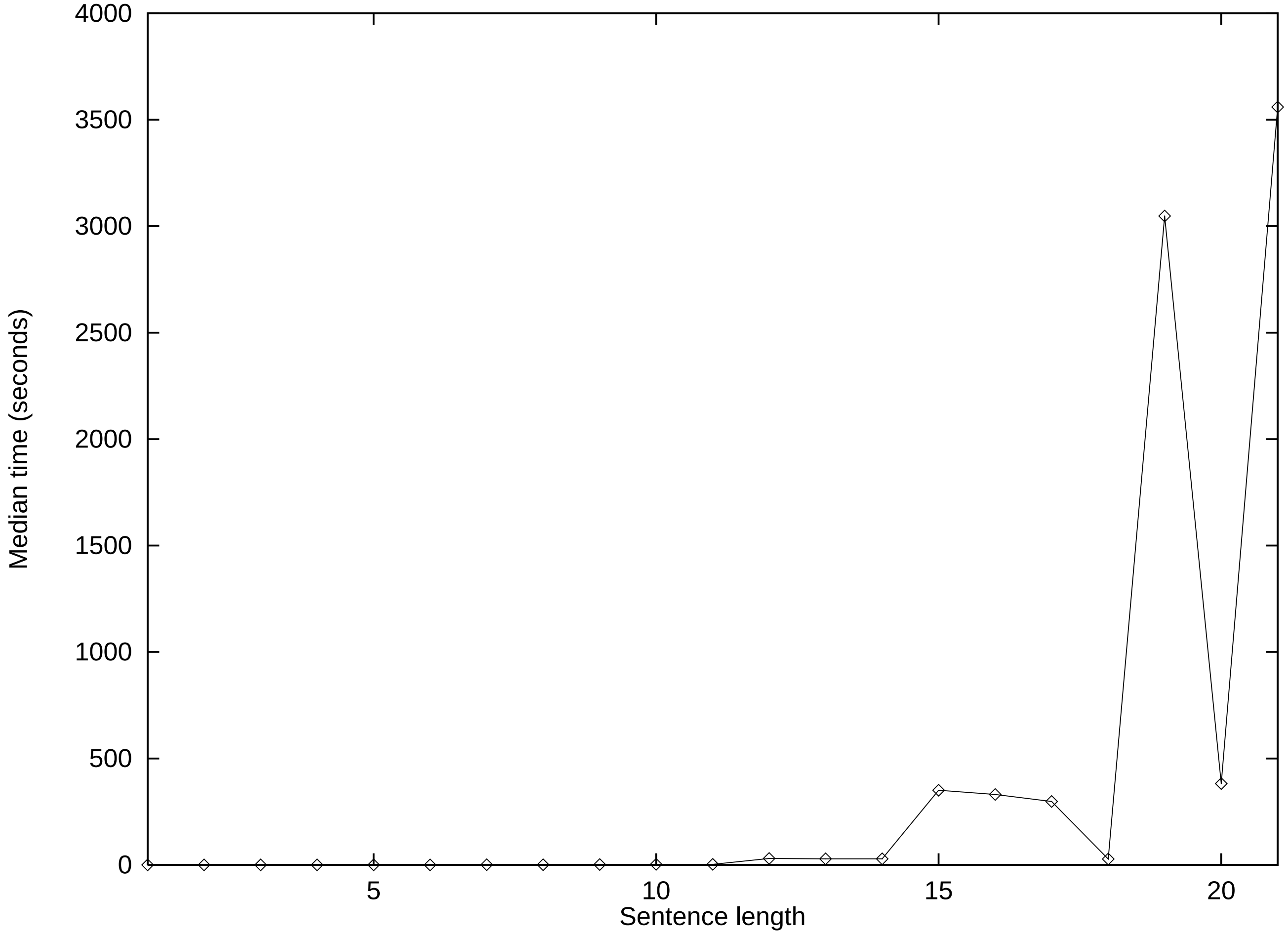


Number of derivations per sentence



Parsing times per sentence

Coeff of determination $R^2 = 0.65$



Median parsing times per sentence

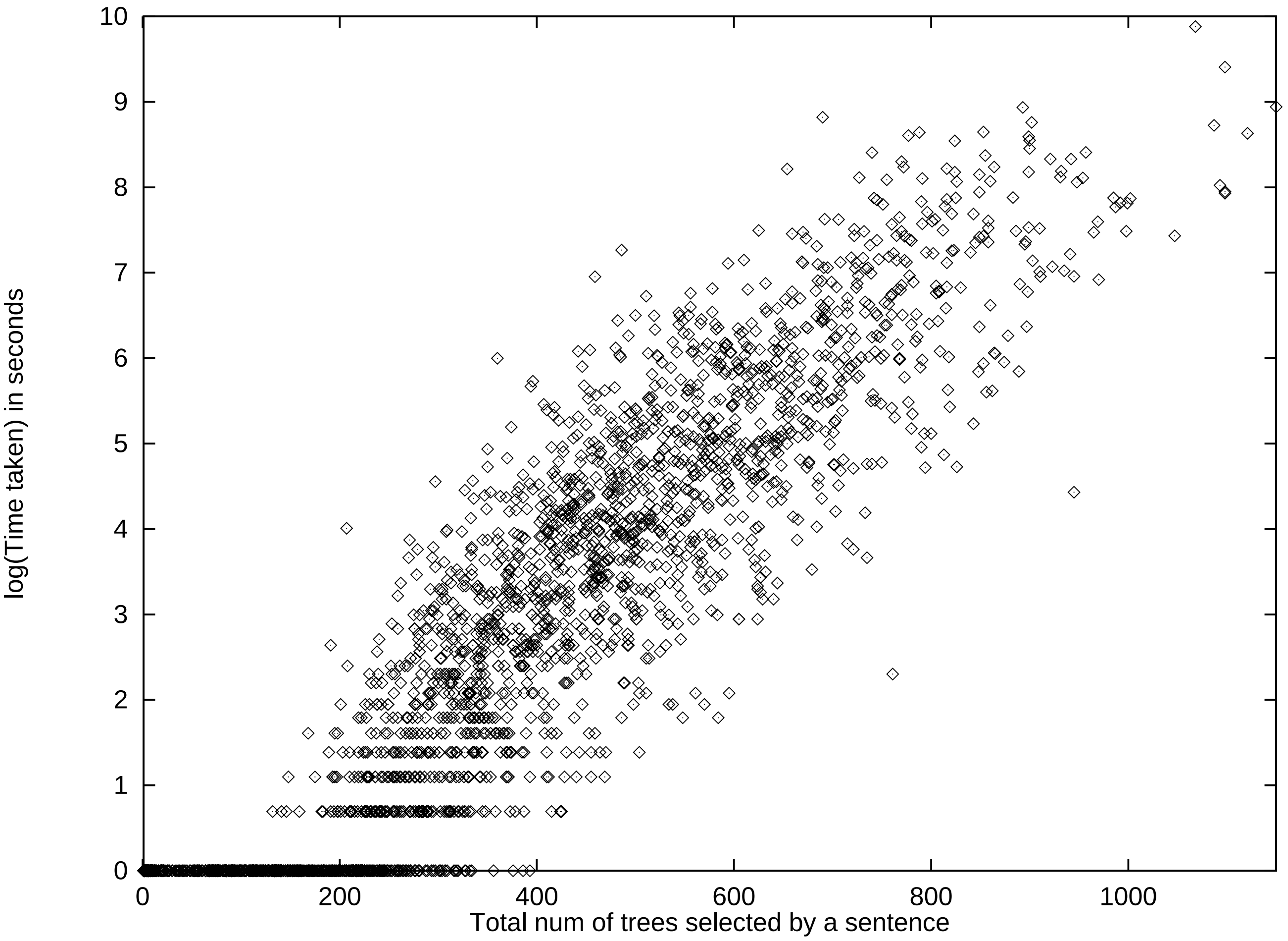
Hypothesis

- There is a large variability in parse times.

- The typical increase in time depending on sentence length is not observed.

- Can a sentence predict its own parsing time?

- Hypothesis: check the number of lexicalized trees that are selected by each sentence.

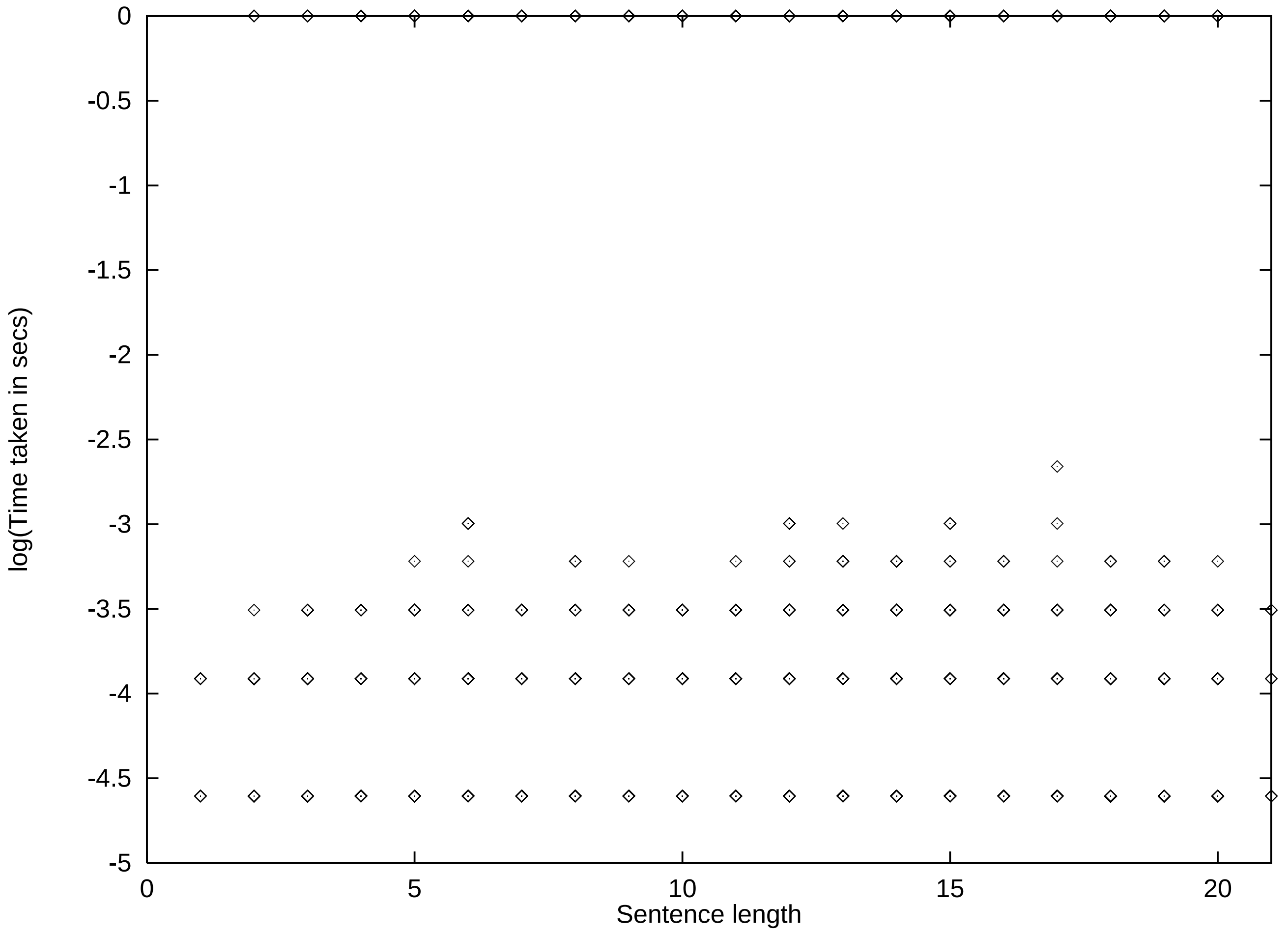


The impact of *syntactic* lexical ambiguity on parsing times

$R^2 = 0.82$ (previous = 0.65)

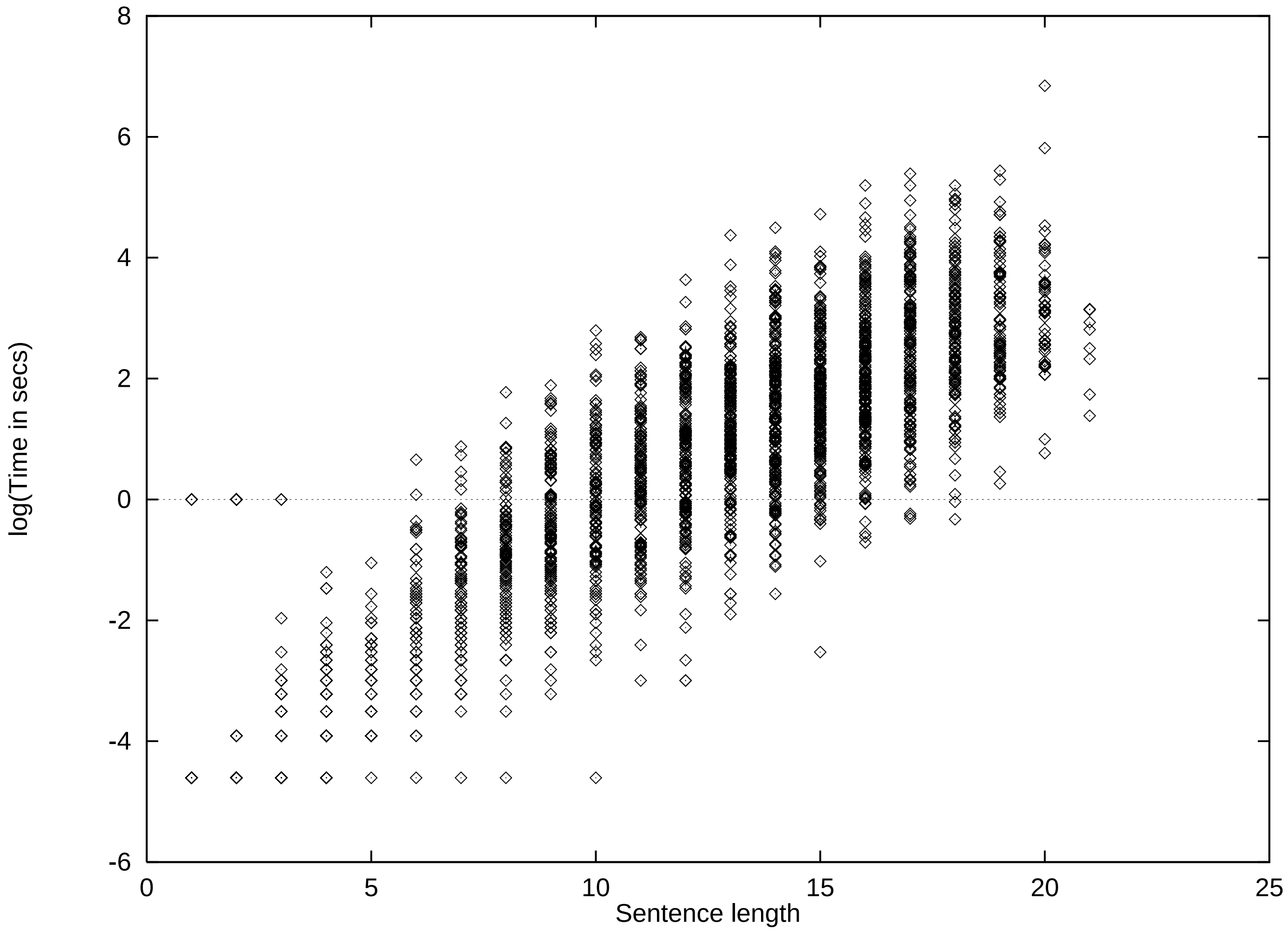
Hypothesis

- To test the hypothesis further we did the following tests:
  - Check time taken when an oracle gives us the single correct tree for each word.
  - Check time taken after parsing based on the output of an n-best SuperTagger.



Parse times when the parser gets the correct tree for each word in the sentence

Total time = 31.2 secs vs. 548K secs (orig)

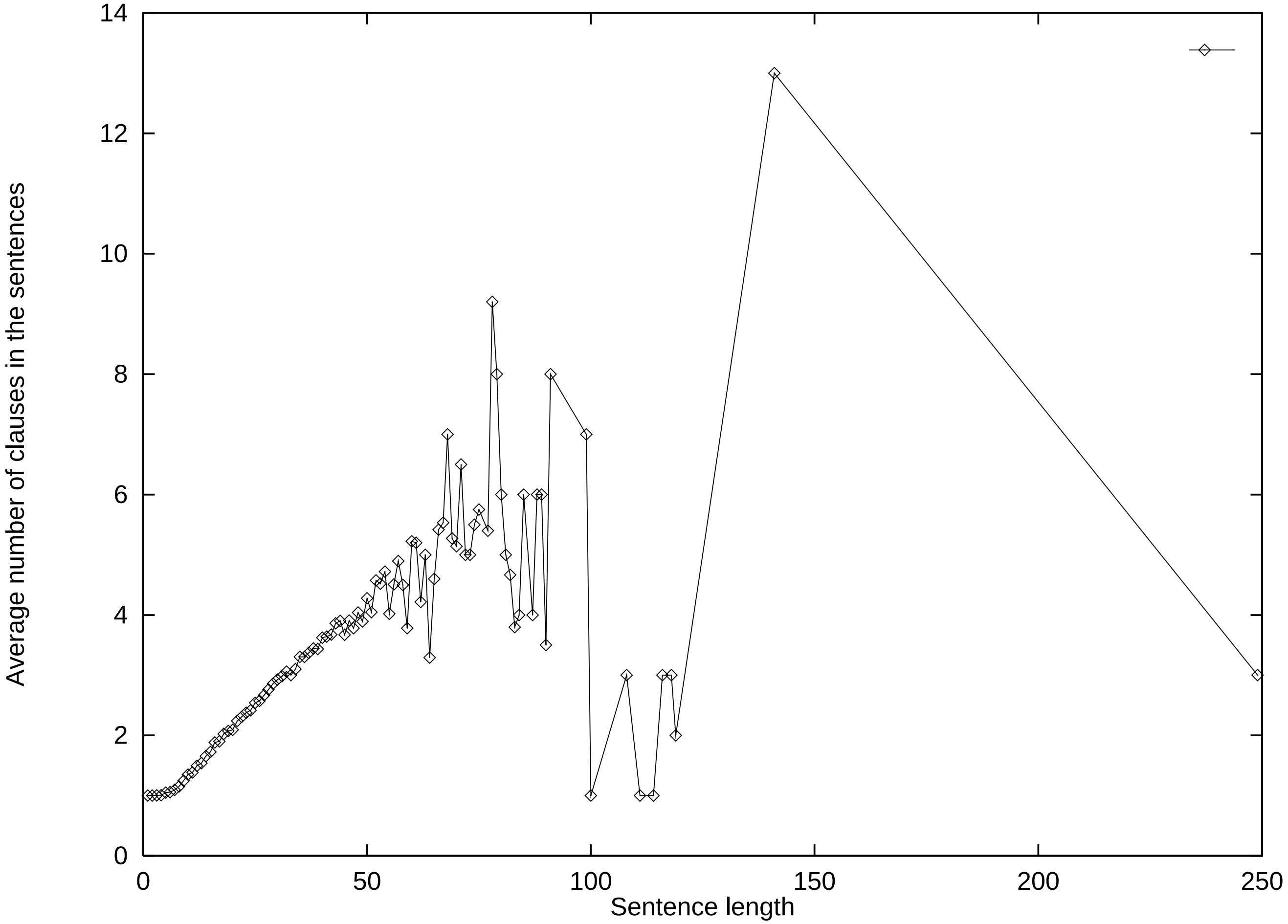


Time taken by the parser after n-best SuperTagging ($n = 60$)

Total time = 21K secs vs. 548K secs (orig)

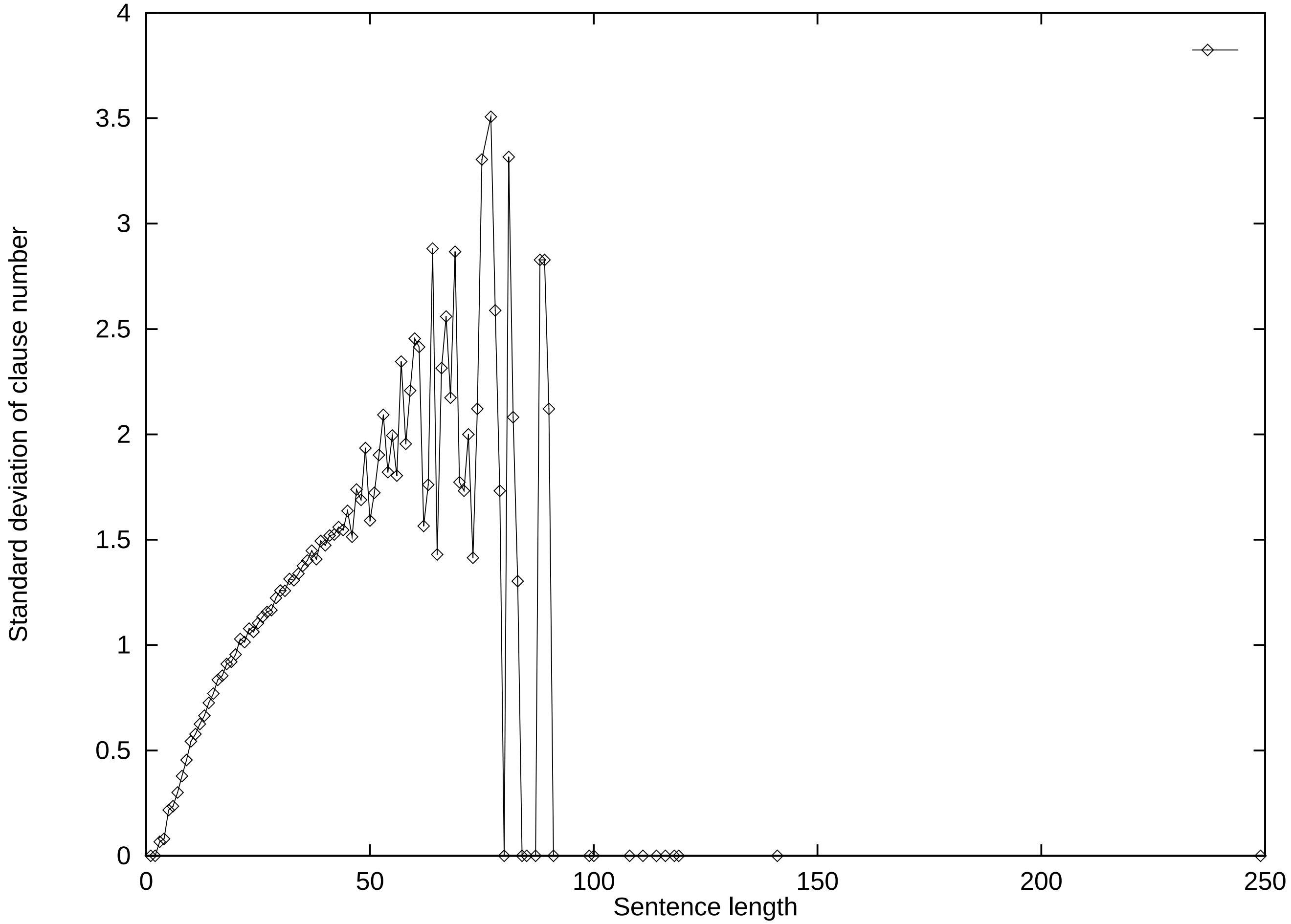
Clausal Complexity

- The complexity of syntactic and semantic processing is related to the number of predicate-argument structures being computed for a given sentence.
- This notion of complexity can be measured using the number of clauses in the sentence.
- Does the number of clauses grow proportionally with sentence length?



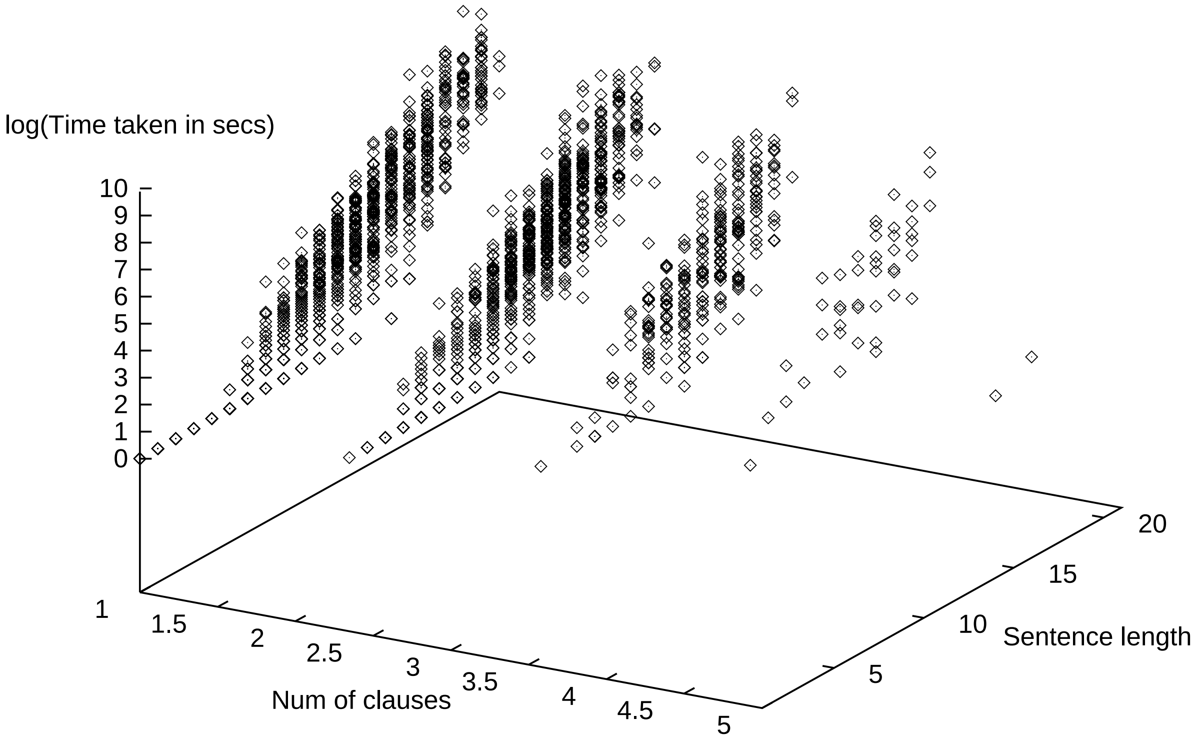
Average number of clause plotted against sentence length.

99.1% of sentences in the Penn Treebank contain 6 or fewer clauses



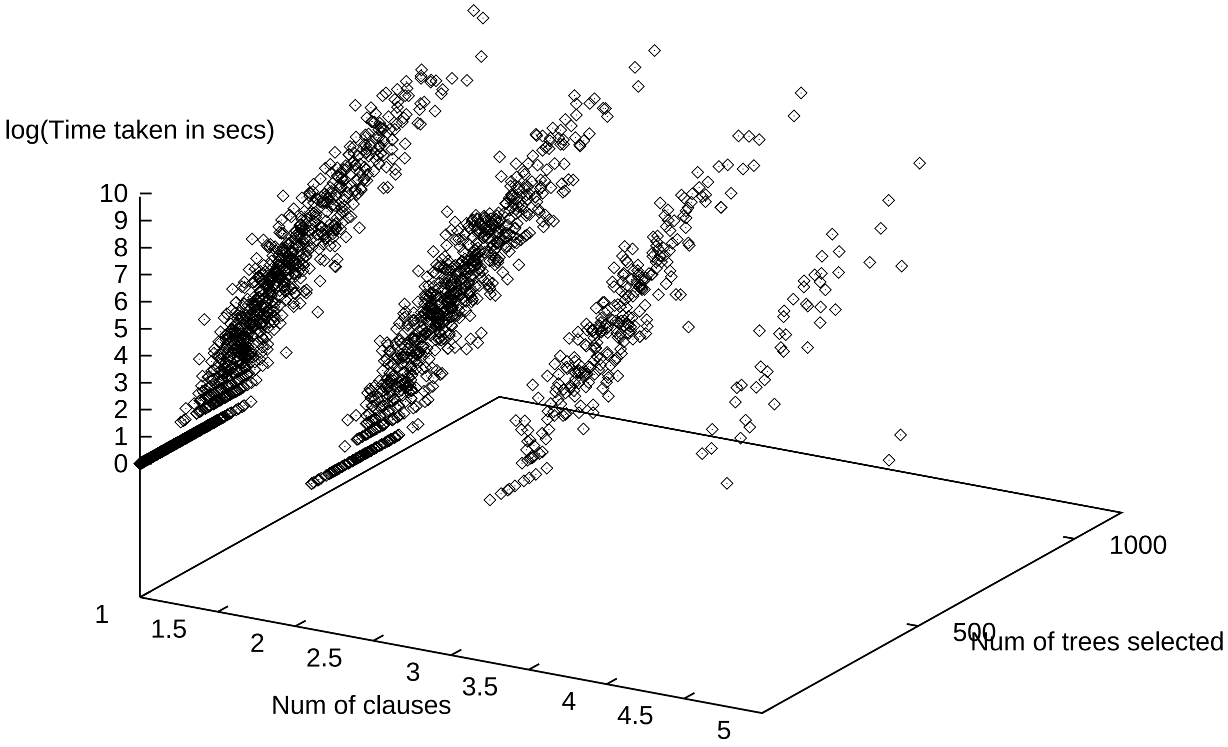
Standard deviation of clause number plotted against sentence length.

Increase in deviation for sentences longer than 50 words.



Variation in parse time against sentence length

while identifying the number of clauses



Variation in parse time against number of trees

The parser spends most of its time attaching modifiers

Conclusions

- We explored two issues that affect parsing effiency for LTAGs: syntactic lexical ambiguity and clausal complexity.
  - Parsing of LTAGs is determined by number of trees selected by a sentence.
  - Number of clauses does not grow proportionally with sentence length.
- Current work: incorporate these factors to improve parsing efficiency for LTAGs.