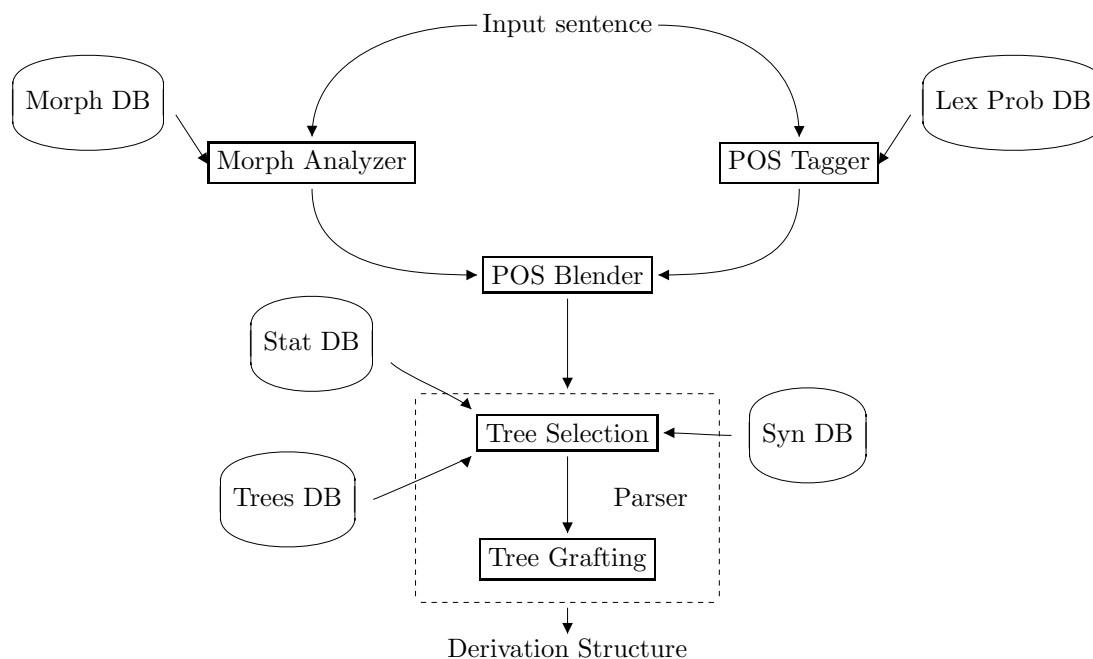


6 The XTAG grammar

XTAG Research Group (2001)

Large coverage LTAG for English. Implemented by the XTAG Research Group at the University of Pennsylvania.

6.1 Architecture of the system



- Morphological Analyzer and Morph Database: Consists of appr. 317000 inflected items derived from over 90000 stems. Returns root form, POS, and inflectional information.
- POS Tagger and Lex Prob Database: Wall-Street Journal trained trigram tagger, extended to output N-best POS sequences.
- Syntactic Database: More than 30000 entries. Each entry consists of: uninflected form of the word, POS, list of trees or tree-families associated with the word, and a list of feature equations that capture lexical idiosyncrasies.
- Tree Database: 1004 trees, divided into 53 tree families and 221 individual trees. The tree families represent subcategorization frames; the trees in a tree-family would be related to each other transformationally in a movement-based approach.
- Tree Selection: For each word, tree templates are chosen from the Tree Database and the anchor position is filled with the word.
- Tree Grafting: once a particular lexicalized tree set is chosen for a sentence, parsing is done. Output: parse tree (derived tree) and derivation tree.

Parsing: TAGs are close to CFG. Therefore algorithms for CFG can be extended to TAG. Schabes and Joshi (1988) describe an Earley-type parser for TAG (top-down parsing on derived tree).¹¹ TAG

¹¹Other references on TAG parsing:

parsing complexity: $\mathcal{O}(n^6)$. Satta (1994) has shown that if there exists an algorithm faster than $\mathcal{O}(n^6)$ for TAG parsing, it is probably not of practical interest because of huge constants.

Supertagging: Before parsing, *supertagging* (Srinivas and Joshi 1999) is possible. This step uses statistical disambiguation to assign a unique elementary tree (a *supertag*) to each word in the sentence.

6.2 Tree Families and Subcategorization Frames

Tree families group together trees belonging to the same subcategorization frame.

E.g., all trees for different forms of *buy* in (38) belong to one tree family. *buy* in (39) has a different tree family.

- (38) a. John bought a book
b. What does John buy?
c. Who bought a book?
d. A book was bought by John
e. The man who bought the book this morning was from Tübingen.

- (39) John bought Mary a book

In each tree family, there is one base tree (e.g., the one for *bought* in (38a.)) and the other trees in the family are obtained from the base tree by transformations (implemented for example by metarules) where the semantic interpretation of arguments remain constant.

Additionally, *lexical rules* transform base trees while changing the properties of particular words within the same subcategorization frame. This leads to a different tree family. E.g., ergative verbs such as *melt* in (40).

- (40) a. The sun melted the ice.
b. The ice melted.

The XTAG grammar covers the following phenomena: auxiliaries, copula, raising and small clause constructions, topicalization, relative clauses, infinitives, gerunds, passives, adjuncts, it-clefts, wh-clefts, PRO constructions, noun-noun modifications, determiner sequences, genitives, negation, noun-verb contractions, clausal adjuncts and imperatives.

Exercise 14 *The simplest verb tree family in XTAG is T_{nx0V} , the tree family for intransitive verbs such as laugh. Find sample sentences containing uses of forms of laugh in different constructions that should be included in this tree family.*

Schabes and Vijay-Shanker (1990); Nederhof (1998) describe an LR parsing algorithm for TAG.

TAG parsing through Boolean matrix multiplication: Rajasekaran and Yooseph (1995).

Boullier (1999, 2000) uses *Range Concatenation Grammars (RCG)* for TAG parsing: The TAG is transformed into an RCG and the RCG is then parsed. The RCG roughly describes the set of derivation trees of the TAG. Allows very fast parsing. Parsing of restricted variant of TAG: Satta and Schuler (1998): not more than two nodes on spines of left/right trees, at most one adjunction of a wrapping tree at the spine of a wrapping tree: $\mathcal{O}(n^5)$.