Efficiently extract recurring tree fragments from large treebanks

Federico Sangati, Willem Zuidema, and Rens Bod

{f.sangati,zuidema,rens.bod}@uva.nl

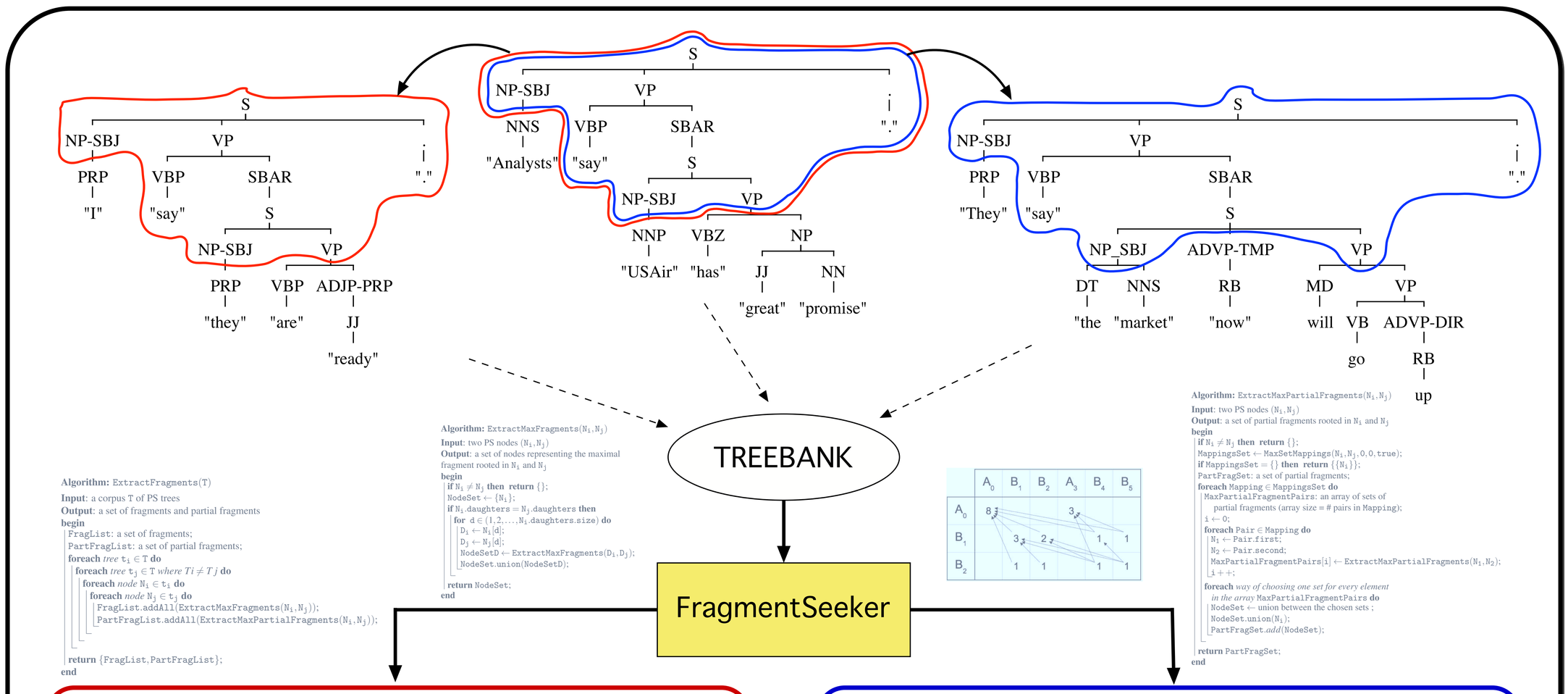
Institute for Logic, Language and Computation

ILLC, University of Amsterdam

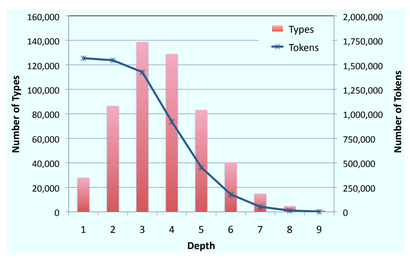
Netherlands Organisation for Scientific Research

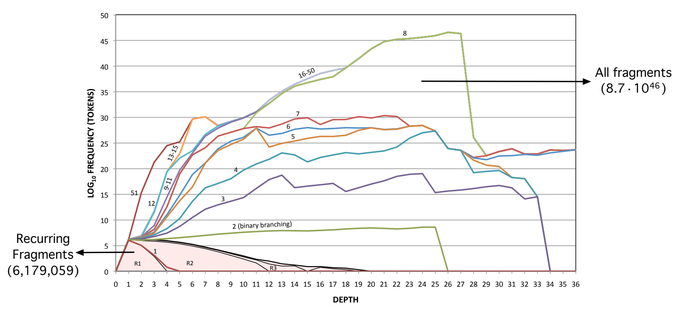
Which are the most relevant syntactic constructions? The ones recurring multiple times!

`FragmentSeeker` **is a tool to extract all tree constructions recurring multiple times in a large treebank**. The tool is based on an efficient kernel-based dynamic algorithm, which compares every pair of trees of a given treebank and computes the list of fragments which they both share. The developed algorithm is much more efficient than the naive approach of extracting all possible fragments from the treebank: their number is extremely large, since it grows exponentially in the size of the treebank trees, but more than 99% occur only once.



WSJ fragment statistcs





Possible Applications

- **Corpus analysis**: valency analysis/extraction.
- **Argument/adjunct distinction**: determine automatically whether a certain node is part of the valency structure of the governing node.
- **Annotation tool**: the availability of the most frequent constructions of a certain lexical item could be beneficial in the annotation and correction process.
- **Parsing**: using the extracted fragments in a Tree Substitution Grammar (TSG) parsing framework.

FragmentSeeker is publicly available at
`http://staff.science.uva.nl/~fsangati`