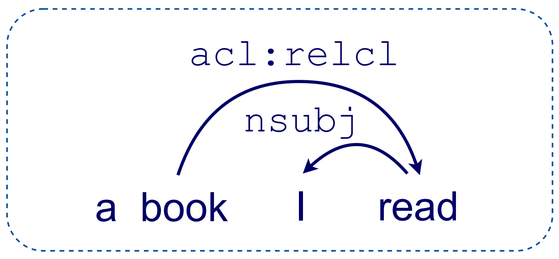
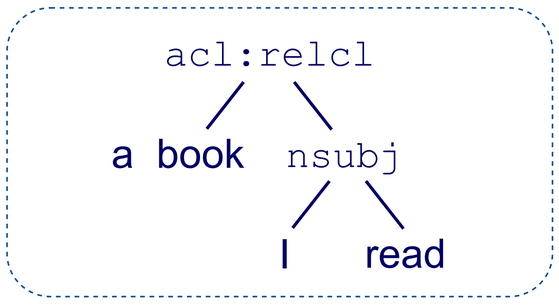
Development of a Multilingual CCG Treebank via Universal Dependencies Conversion

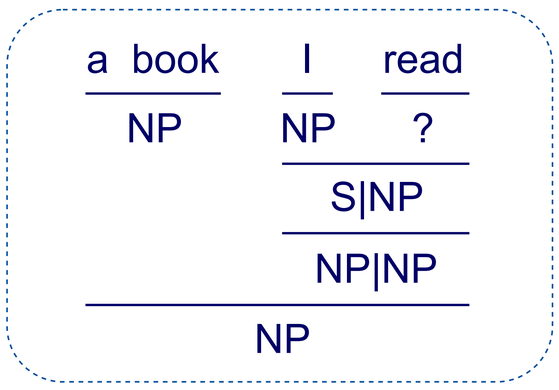
Tu-Anh Tran and Yusuke Miyao
The University of Tokyo

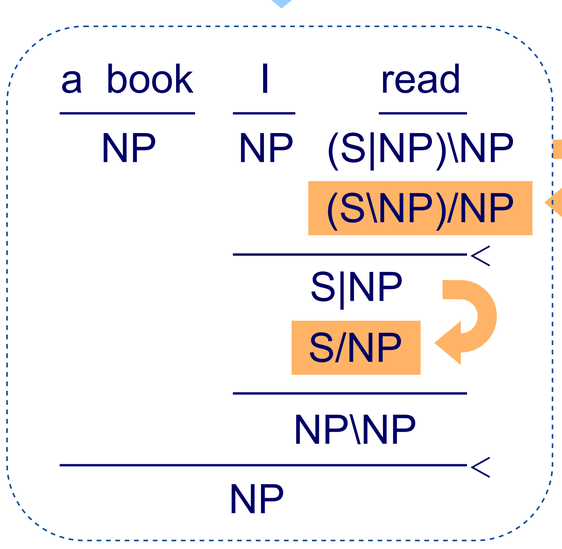
Summary

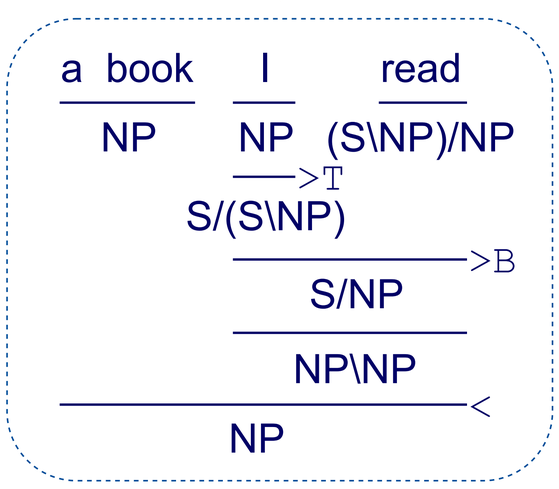
We propose a rule-based algorithm to create CCG treebanks from UD treebanks for many languages.











Binarization

Binarize dependency trees based on a pre-defined obliqueness hierarchy to match the binary structures of CCG derivations.

Category Assignment

Apply hand-crafted rules that assign CCG categories to noun phrases, modifiers, function words, and punctuation marks.

Category Inference

Apply CCG's combinatory rules to infer the categories of unassigned constituents.

Phase 1

Category Reassignment via Majority Voting

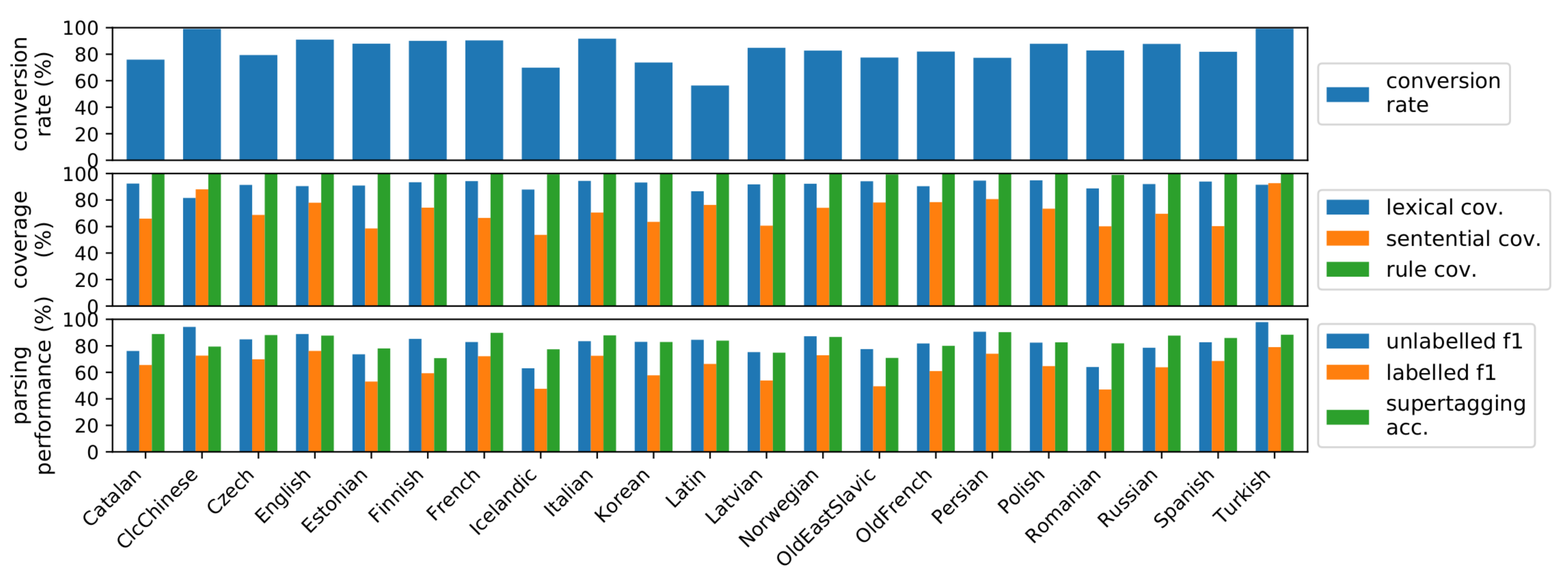
Gather votes from assigned categories in Phase 1 to decide the most common word order of a treebank. Based on the most common word order, reassign CCG categories to predicates.

Parsing and Filtering

Apply a non-statistical CCG parser to generate candidate parses given currently assigned categories. Filter out parses that do not match our binarized trees and hand-crafted rules in Phase 1.

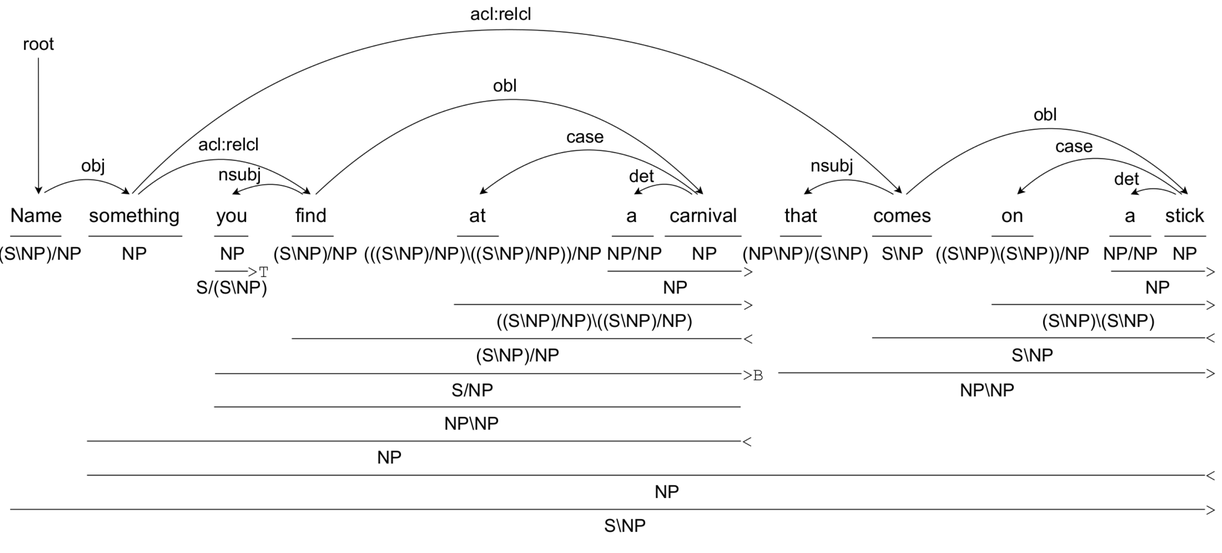
Phase 2

Results



Coverage and parsing performance on select 21 CCG treebanks of 21 languages.

Example with Long-distance Dependency



(English_EWT: `sent_id = answers-20111107163942AA08rP5_ans-0001`)

Conclusion

- Limitations:
  - Languages with flexible word order
  - Long-distance dependencies with no explicit traces
  - Argument/adjunct distinction for PP
  - Crossing dependencies
- Ideas for the future:
  - Incorporating Enhanced Dependencies
  - Incorporating Universal Proposition Banks*

* https://universalpropositions.github.io/