Steps in the direction of testing a GF RG

Herbert Lange
University of Gothenburg

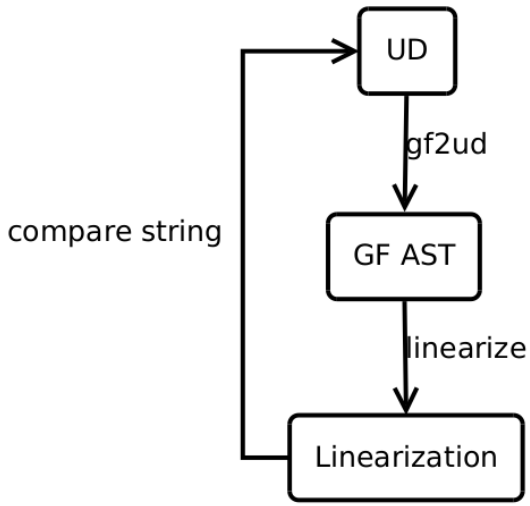
August 24, 2017

Two approaches

- Top-down
- Bottom-Up

Top-Down

Workflow:



Problems: Missing Constructions in the grammar, variations in linearization

Bottom-Up

- Extend lexicon
- Test against state-of-the-art morphology
- Test coverage against treebank
- Test sentence coverage

Extend lexicon

- Whitaker's Words (`http://archives.nd.edu/whitaker/wordsdoc.htm`) - 39225 Entries
- Perl script and manual work (until letter "d", 45%) $\Rightarrow$ 31507 concrete Entries out of 37135 abstract ones

Problems: Abbreviations, undeclinable proper names, comparison levels for adverbs, and Greek adjectives

Test against state-of-the-art morphology

- LatMor by Uwe Springman (`http://www.cis.uni-muenchen.de/~schmid/tools/LatMor/`, based on SFST)
- Generated 2184 noun forms, 3740 verb forms, and 5184 adjective forms with GF and analyzed them with LatMor
- Recognized 2095 noun forms (89 unknown, 96% recognized), 2282 verb forms (1458 unknown, 61% recognized) and 4731 adjective forms (453 missing, 91% recognized)

Problems: Modern words for LatMor

Test coverage against treebank

- Caesar's "Commentarii de Bello Gallico" in UD: 1329 Sentences, 6600 unique Tokens, 2491 unique Lemmas
- LatMor recognizes 6520 Tokens (80 unknown, 99% recognized) and 2434 Lemmas (57 unknown, 98% recognized)
- GF recognizes 3935 Tokens (2585 unknown, 60% recognized) and 1727 Lemmas (764 unknown, 69% recognized)

Problems: Normalization and lexicon coverage (for GF)

Test sentence coverage

Maybe next summer school