Linking the ERG lexicon and English WordNet

Dan Flickinger

CSLI, Stanford University

DELPH-IN Summit 2022

Fairhaven/Bellingham

18 July 2022

Aim: full ERG lexical coverage of WordNet entries

- Do better linguistics for the core and the periphery
  - Lexical idiosyncracy, multi-word expressions
- Reduce limitations on generation
  - Difficult to predict lexical type from predicate
  - Stemming ambiguous: *_devined/VBD_u_unknown*
- Reduce dependence on imperfect POS-taggers for parsing
  - 97% word-accuracy: 56% sentence-accuracy (Manning:2011)
  - With 44K manual lexicon, still 9500 unknowns in WSJ
- Provide stable basis for word-sense tagging of corpora
  - *_cleverness/NN_u_unknown* vs *_ingenuity_n_1*

Challenges

- Size: 155K words in 176K synsets, 207K word-sense pairs
- Mismatches between WordNet entries and lexemes
- Exclusion of proper names? If so, what criteria?
- Naming semantic predicates: grammatically distinct senses

Possible risks

- Some processing engines/tools might balk at 150K lexicon
- Parse selection model may do worse on (rarer) unknown words