Linguistic Core Year 3: Resources

Noah Smith

CMU

Where to Find Everything

Listing with links:

`https://sites.google.com/site/linguisticcore/developed-resources`

New Data Resources

- Parallel
  - Swahili-English (1M words)
- High-quality parallel
  - Swahili-English (130K words)
  - Malagasy-English (410K words)
- Monolingual
  - Kinyarwanda (6M words)
- **Annotated**
  - Part-of-speech: types and tokens for Kinyarwanda, Malagasy (5K types, 10K tokens, each)
  - Graph Fragment Language: Kinyarwanda (4K sentences), Malagasy (8K sentences), English (1K sentences)
  - Definiteness: English, Russian, Hindi (18K words, $\kappa > 0.9$)

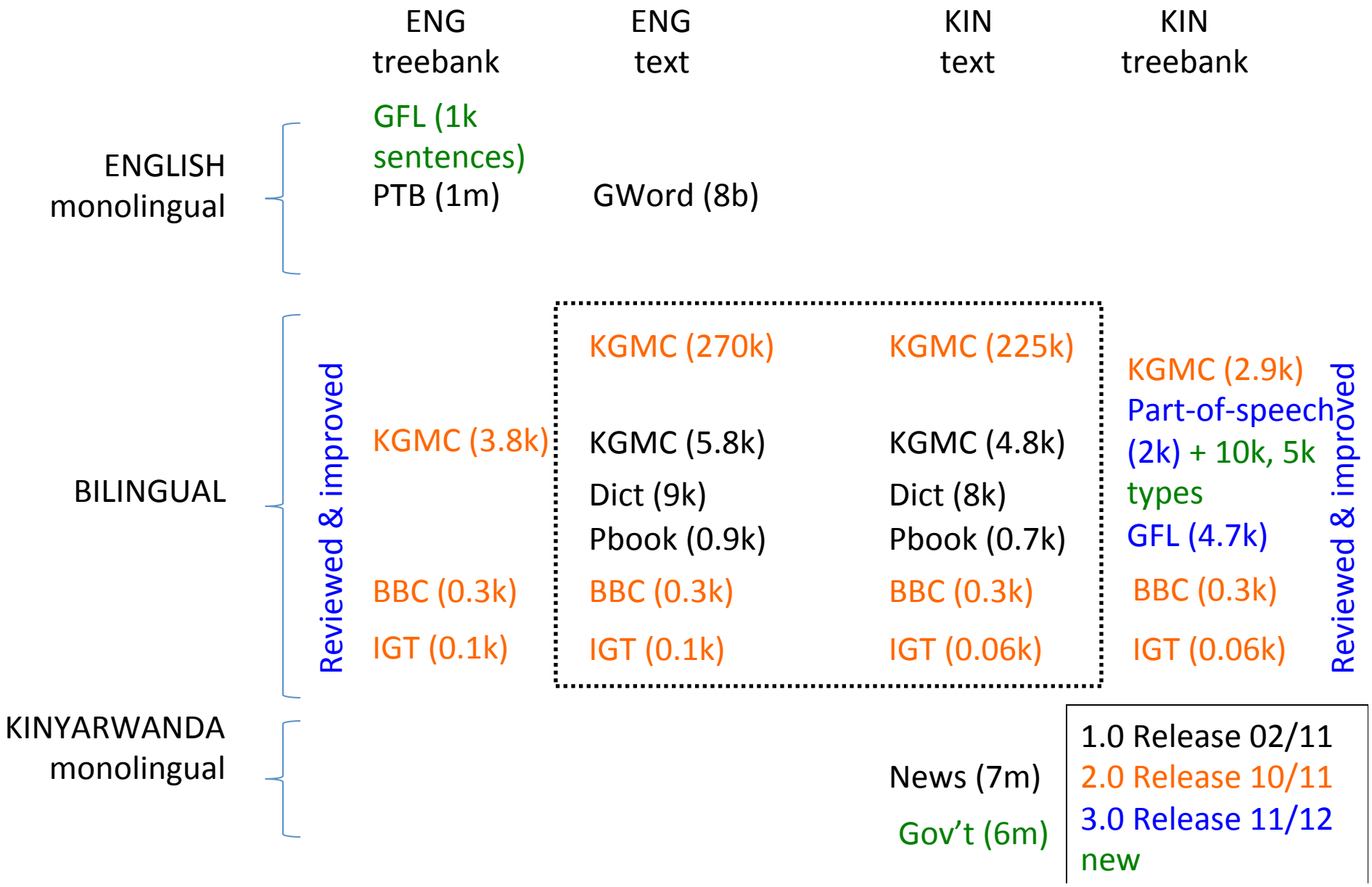
Who will tell us more:

Dan Garette

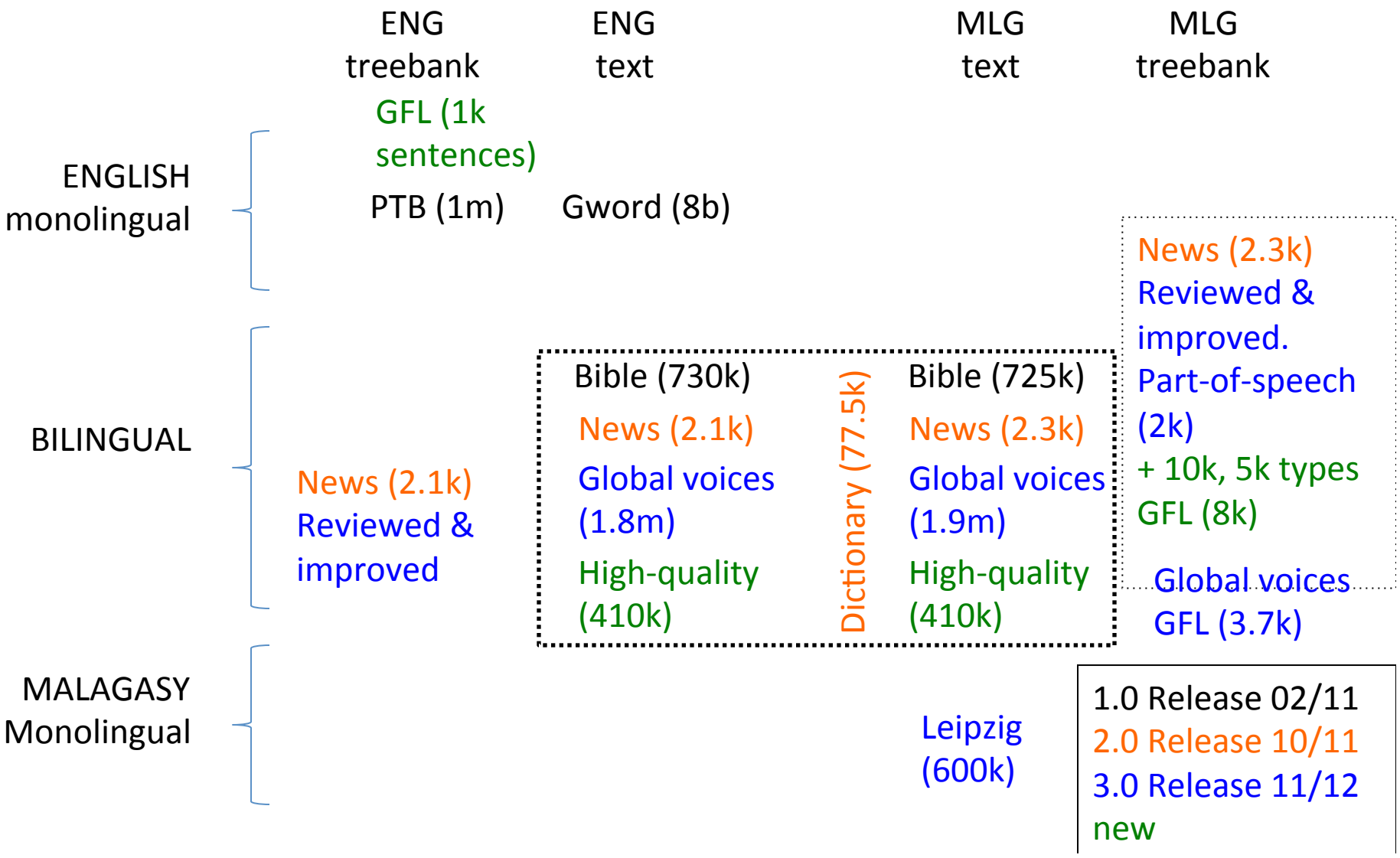
Jason Baldridge

Lori Levin

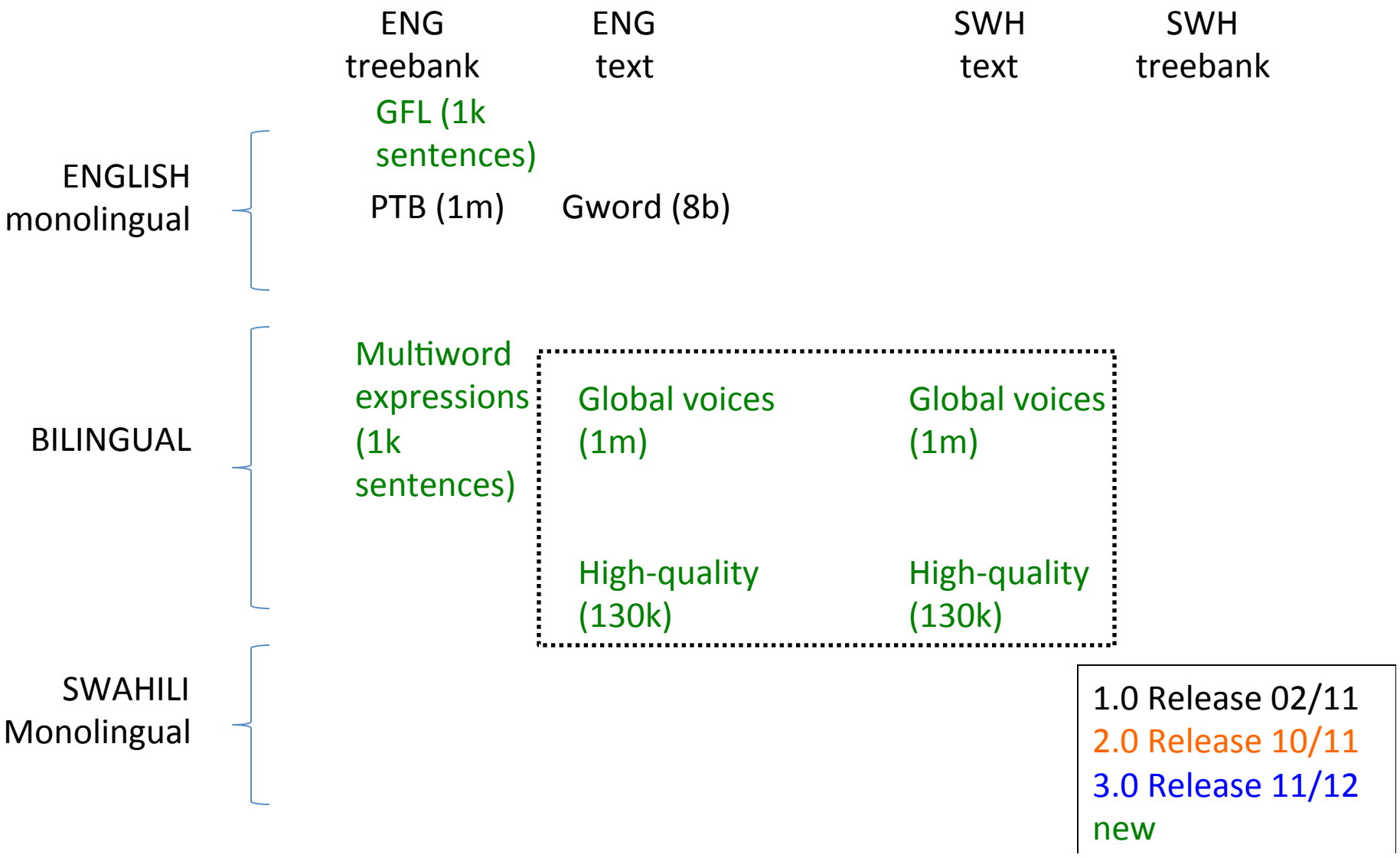
Kinyarwanda Data Resources



Malagasy Data Resources



Swahili Data Resources



New Tools: Text Analysis

- Morphological analyzers
  - Finite-state transducer lexicons: Kinyarwanda, Swahili
  - `fast_umorph`: morphological grammar induction
- Part-of-speech tagging
  - `TurboTagger` for Kinyarwanda, Malagasy
  - Low-resource tools
- Dependency parsing
  - `TurboParser` and new parser trained for Kinyarwanda, Malagasy
  - GFL tools for syntactic annotation, verification, visualization

New Tools: Translation

- `Bolinas`: synchronous hyperedge replacement grammars/algorithms
- `fast_align`: word alignment
- `morphogen`: translation with synthetic phrases
- `morpholm`: morphological language modeling
- `hols`: machine translation parameter estimation

New Data & Resources: Summary

- *New data:* parallel, "high-quality" parallel, monolingual, syntactic and semantic annotation
- *New in TA:* FSTs, morph. grammar induction, POS, dependency parsers, annotation tools
- *New in MT:* SHRGs, word alignment, morphological language modeling, synthetic phrase modeling, parameter estimation