

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 Dan Garette
(5K types, 10K tokens, each)
 - Graph Fragment Language: Kinyarwanda (4K sentences), Jason Baldridge
Malagasy (8K sentences), English (1K sentences)
 - Definiteness: English, Russian, Hindi (18K words, $\kappa > 0.9$) Lori Levin

Who will tell us more:

Kinyarwanda Data Resources

	ENG treebank	ENG text	KIN text	KIN treebank
ENGLISH monolingual	GFL (1k sentences) PTB (1m)	GWord (8b)		
BILINGUAL	KGMC (3.8k)	KGMC (270k)	KGMC (225k)	KGMC (2.9k)
		KGMC (5.8k)	KGMC (4.8k)	Part-of-speech (2k) + 10k, 5k types
		Dict (9k)	Dict (8k)	GFL (4.7k)
		Pbook (0.9k)	Pbook (0.7k)	BBC (0.3k)
		BBC (0.3k)	BBC (0.3k)	IGT (0.06k)
IGT (0.1k)	IGT (0.1k)	IGT (0.06k)	IGT (0.06k)	
KINYARWANDA monolingual			News (7m) Gov't (6m)	1.0 Release 02/11 2.0 Release 10/11 3.0 Release 11/12 new

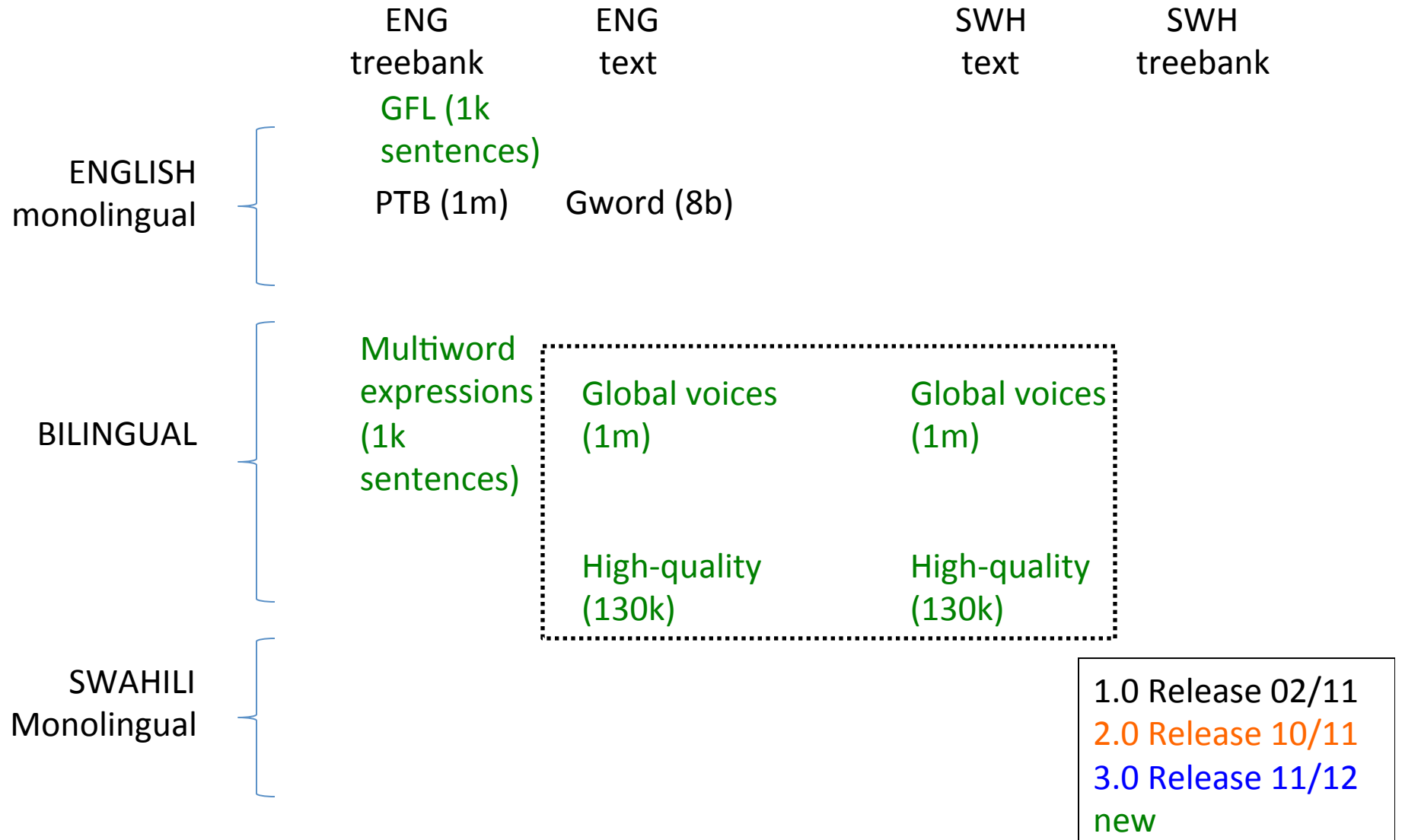
Reviewed & improved

Reviewed & improved

Malagasy Data Resources

	ENG treebank	ENG text	MLG text	MLG treebank
ENGLISH monolingual	GFL (1k sentences) PTB (1m)	Gword (8b)		
BILINGUAL	News (2.1k) Reviewed & improved	Bible (730k) News (2.1k) Global voices (1.8m) High-quality (410k)	Dictionary (77.5k) Bible (725k) News (2.3k) Global voices (1.9m) High-quality (410k)	News (2.3k) Reviewed & improved. Part-of-speech (2k) + 10k, 5k types GFL (8k) Global voices GFL (3.7k)
MALAGASY Monolingual			Leipzig (600k)	1.0 Release 02/11 2.0 Release 10/11 3.0 Release 11/12 new

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
- `morpho1m`: 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