

LDMT MURI

# Data and Technology Transfer

November 14, 2014

Chris Dyer, Carnegie Mellon



Language  
Technologies  
Institute

# What can we do for you?

- **Data**

- Kinyarwanda
- Malagasy
- Swahili
- Yoruba

- **Code**

- Data gathering
- Annotation
- Linguistic analysis
- Modeling

# Kinyarwanda Data Resources

word counts	ENG treebank	ENG text	KIN text	KIN treebank
ENGLISH monolingual (huge)	PTB (1m)	GWord (8b)		
BILINGUAL (285k)	KGMC (3.8k)	KGMC (270k)	KGMC (225k)	KGMC (2.9k)
		KGMC (5.8k)	KGMC (4.8k)	Part-of-speech (2k)
		Dict (9k)	Dict (8k)	GFL (4.7k)
		Pbook (0.9k)	Pbook (0.7k)	
		BBC (0.3k)	BBC (0.3k)	BBC (0.3k)
IGT (0.1k)	IGT (0.1k)	IGT (0.06k)	IGT (0.06k)	
KINYARWANDA monolingual (7m)			News (7m)	1.0 Release 02/11 2.0 Release 10/11 3.0 Release 03/13

Reviewed & improved

R

# Malagasy Data Resources

	ENG treebank	ENG text	MLG text	MLG treebank
ENGLISH monolingual (huge)	PTB (1m)	Gword (8b)		
BILINGUAL (732k)	News (2.1k) Reviewed & improved.	Bible (730k) News (2.1k) Global voices (3.0m)	Dictionary (77.5k) Bible (725k) News (2.3k) Global voices (2.5m)	News (2.3k) Reviewed & improved. Part-of-speech (2k) Global voices GFL (3.7k)
MALAGASY Monolingual			Leipzig (600k)	1.0 Release 02/11 2.0 Release 10/11 3.0 Release 03/13 4.0 Release 06/14

# New mg data from Qing Dao

- 15.3M tokens of Malagasy crawled from news sites
- 396M tokens of web English news about Africa

# Release 4.0 Data

## Data (Swahili ↔ English)

	Tokens	Types	Hapax	Non-language
Parallel [Swahili]	303,941	32,554	20,518	4,721
Parallel [English]	301,148	20,813	9,780	4,899

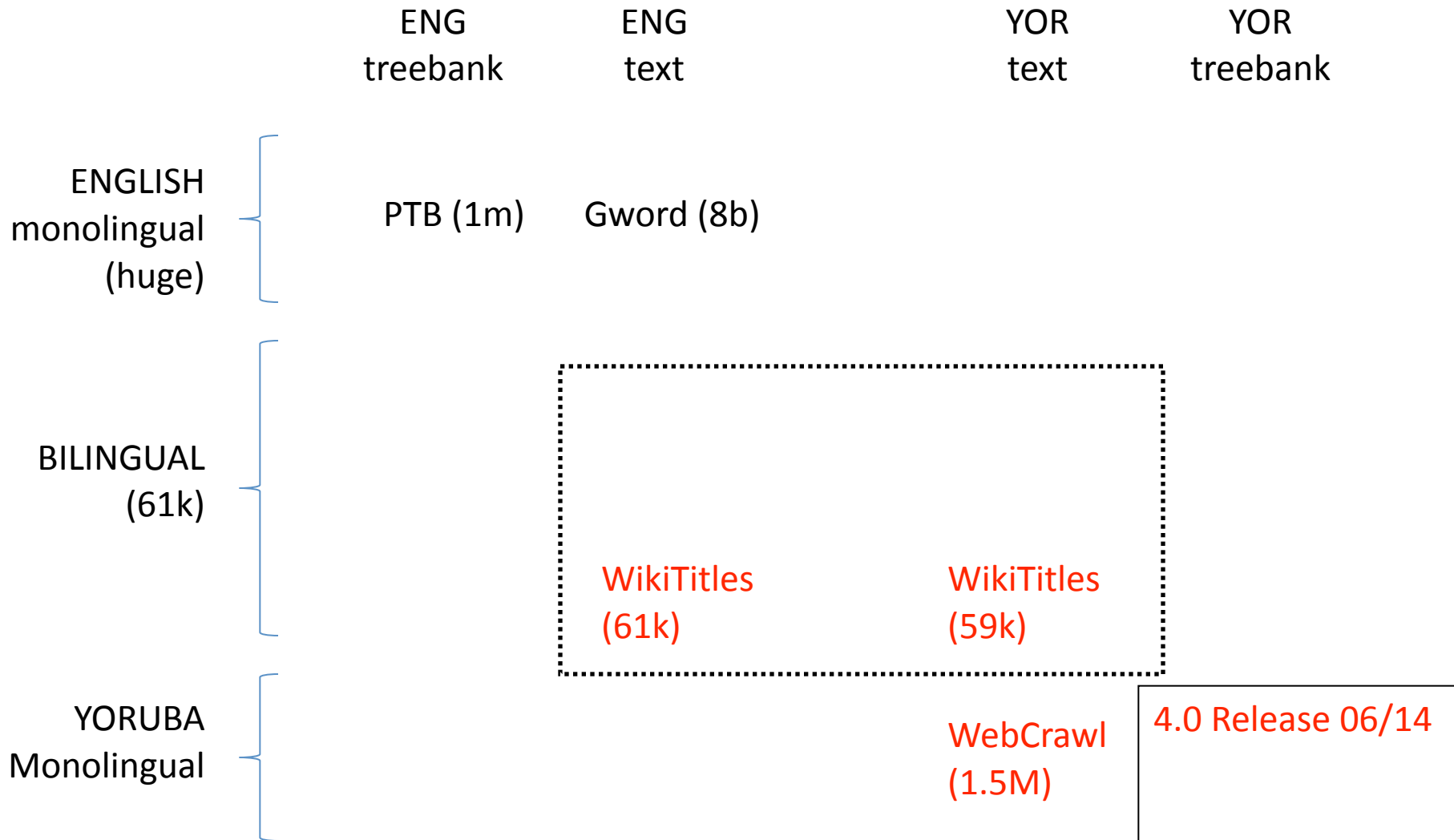
## Data (Swahili only)

	Tokens	Types	Hapax	Non-language
Helsinki Swahili Corpus*	12,534,468	309,615	162,798	168,941

## Data (Malagasy ↔ English)

	eng-Tokens	eng-Types	mlg-Tokens	mlg-Types
Bible (Year 1)	584,872	13,084	579,578	19,460
CMU Global Voices (Year 3)	2,351,923	74,330	2,830,060	97,546

# Yoruba Data Resources



# Technology Releases

- Data gathering
- Data annotation
- Linguistic analysis
  - syntax
  - semantics
- Statistical modeling
  - language
  - translation

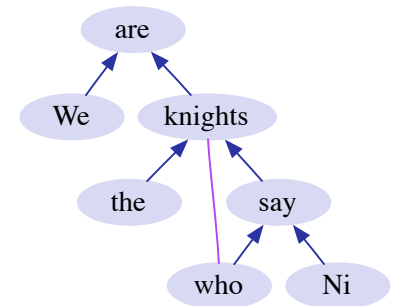
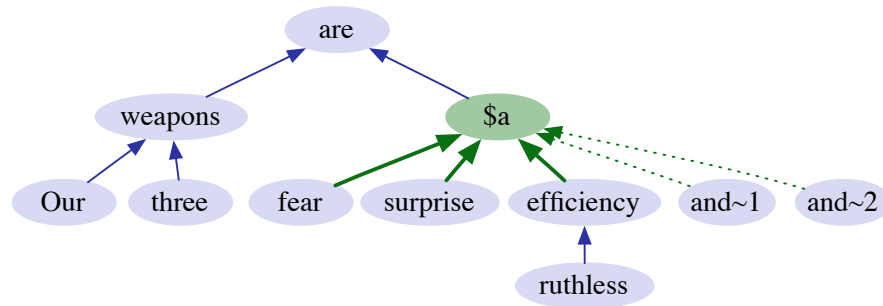
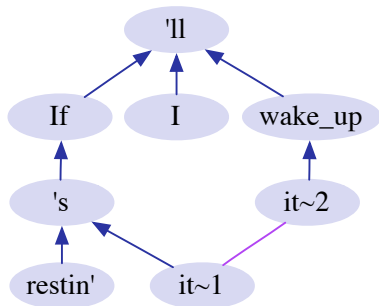


# Data Crawling

- Parallel data crawlers
  - WikiTitles
    - 206 languages currently available
  - GlobalVoices
    - 18 languages currently available
- Outputs
  - Word sentence aligned parallel data
  - Metadata about documents
  - Incremental updates

# Annotation

- GFL Parser
- GFL Annotation Web Interface



- Represents unlabeled dependencies
- Special handling for:
  - multiword expressions
  - coordination
  - anaphora
- Allows underspecification
- Graph fragment language for easy annotation

# Statistical Modeling Toolkits

- **Language Modeling**

- morpholm (rule-based morphology + statistics)
- spectral-lm - low dimensional LM

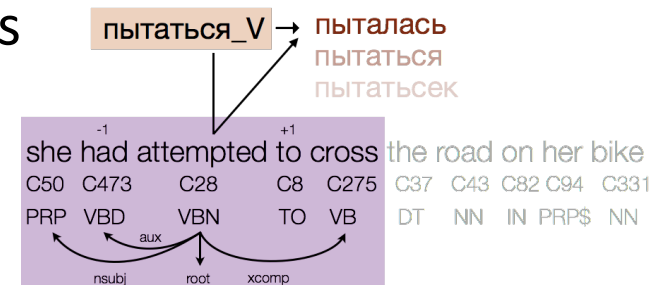
	modint-KN	PLRE
English-Large	77.90 +/- 0.20	<b>75.66 +/- 0.19</b>
Russian-Large	289.6 +/- 6.82	<b>264.59 +/- 5.84</b>

- **Translation Modeling**

- fast\_align - fast word alignment
- spectral-scfg - fast grammar refinement for MT
- morphogen - generate synthetic phrases

- **Forthcoming**

- unsupervised CRF autoencoder suite
- decipherment tools

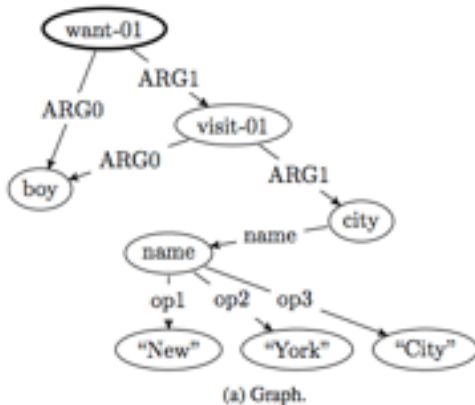


# Linguistic Analysis Toolkits

- Syntax
  - RGBParser
    - fast, state-of-the-art dependency parser
  - CCG Supertagger
    - train using a mixture of token- and type-supervision
  - CCG Parser [next year]
    - train using partial GFL supervision

# Linguistic Analysis Toolkits

- Semantics
  - JAMR - graph-based parser for AMR
  - Bolinas - AMR with graph transducers



- yMetaphor - detect figurative language

# Where?

<http://www.linguisticcore.info>