LDMT MURI

Data and Technology Transfer

November 14, 2014

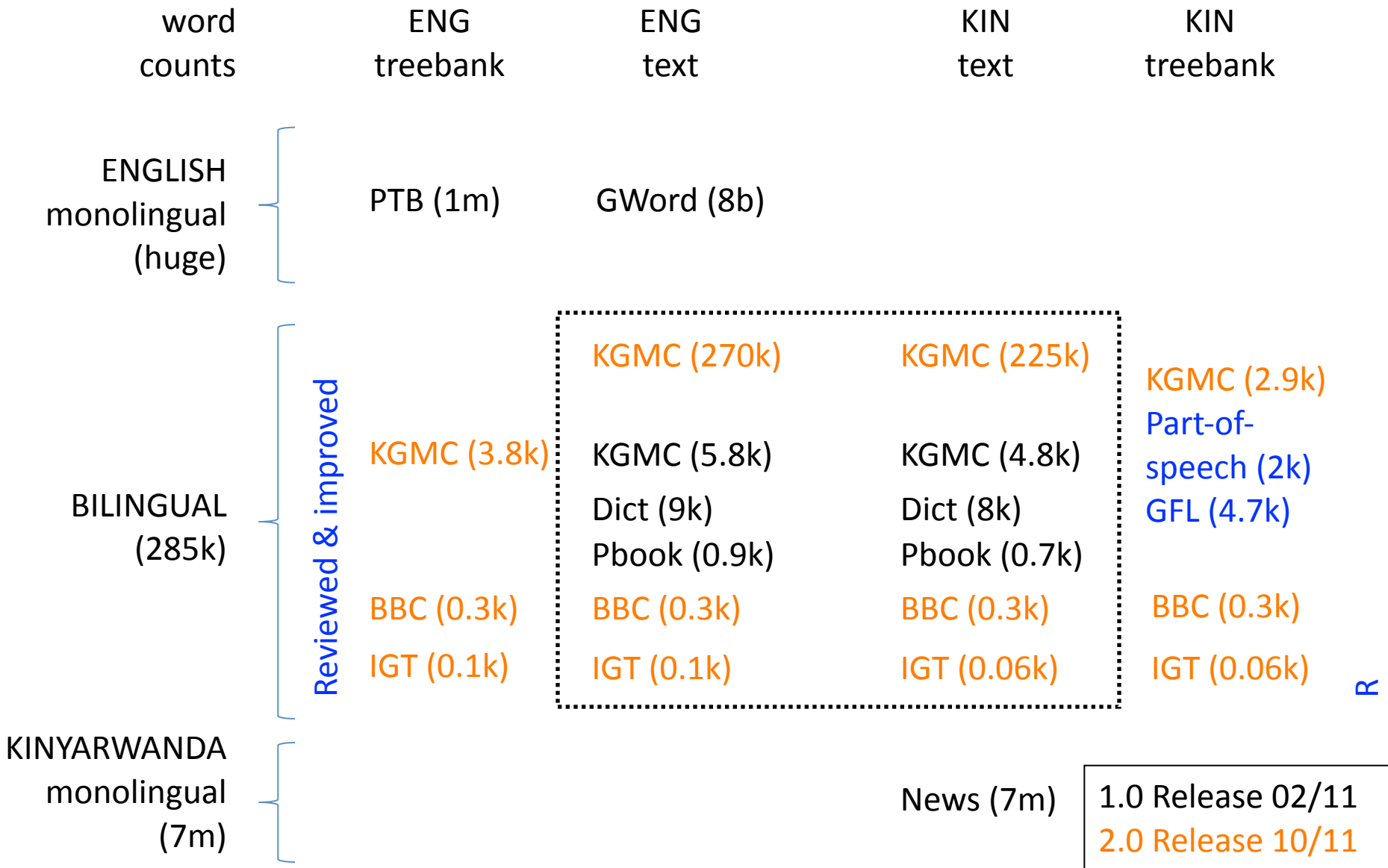
Chris Dyer, Carnegie Mellon

Language Technologies Institute

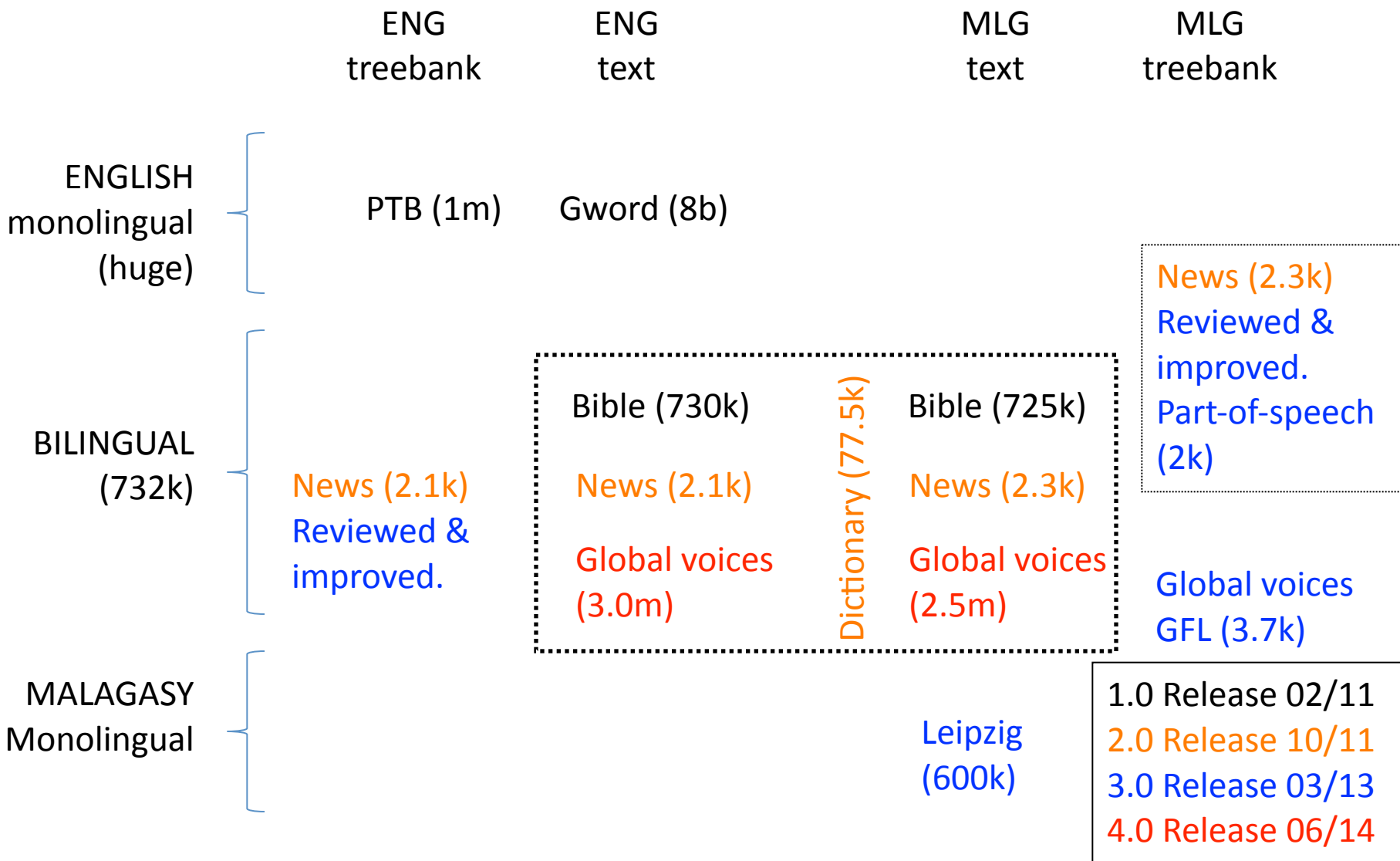
What can we do for you?

- **Data**
  - Kinyarwanda
  - Malagasy
  - Swahili
  - Yoruba
- **Code**
  - Data gathering
  - Annotation
  - Linguistic analysis
  - Modeling

Kinyarwanda Data Resources



Malagasy Data Resources



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

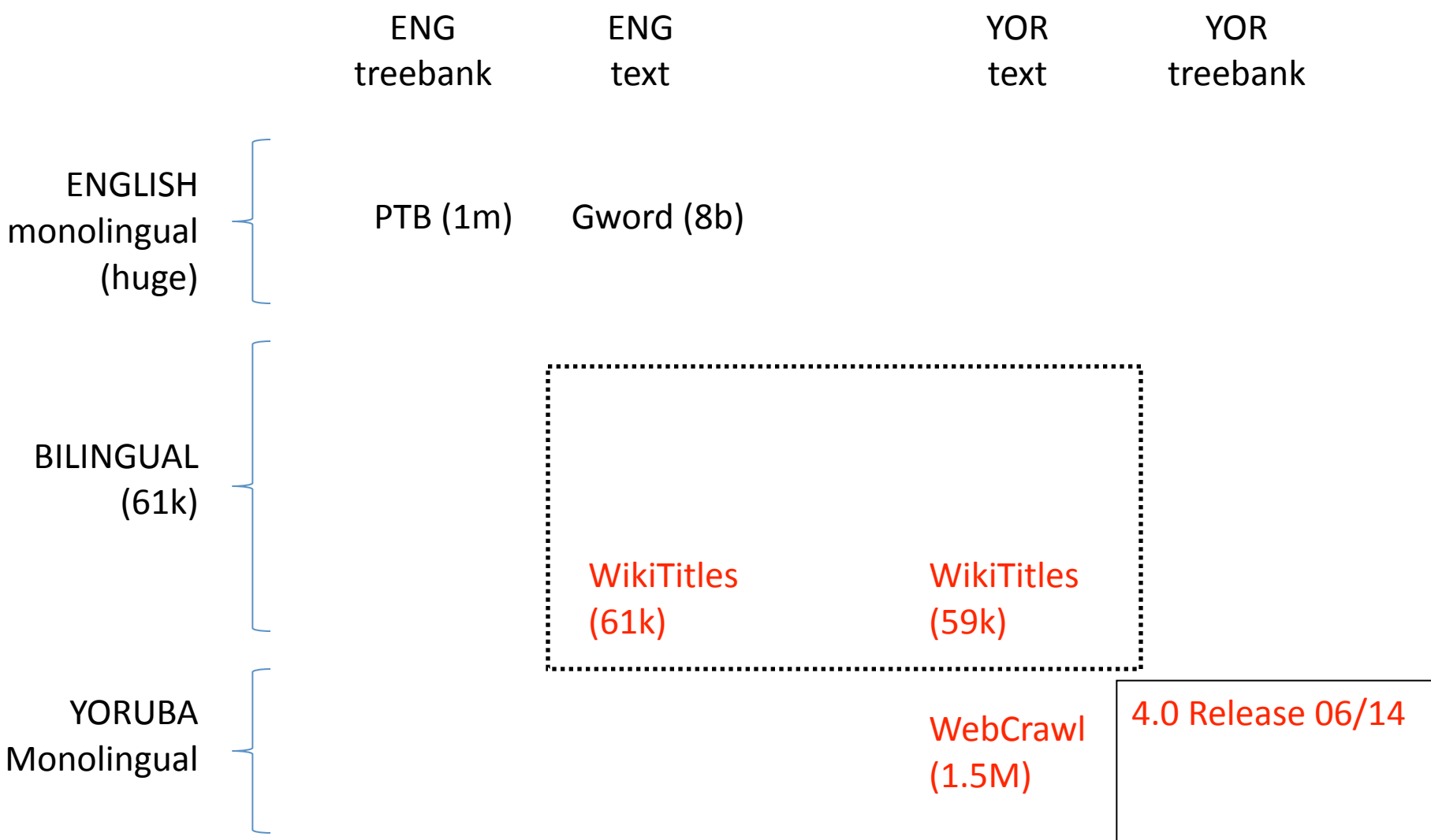
Data (Swahili only)

	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

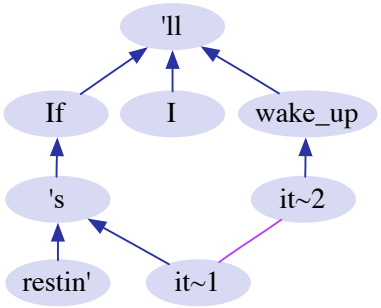
Data Crawling

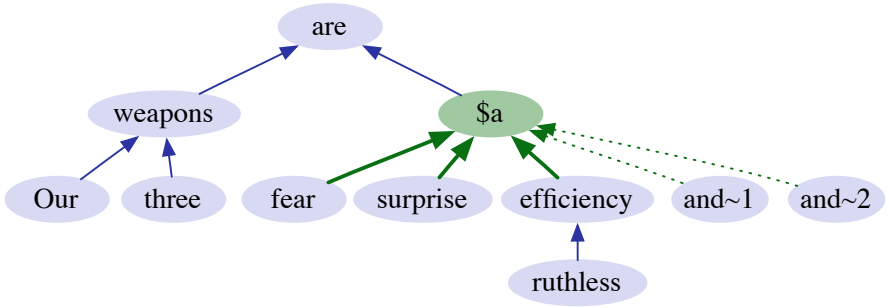
- Parallel data crawlers
  - WikiTitles
    - 206 languages currently available
  - GlobalVoices
    - 18 languages currently available

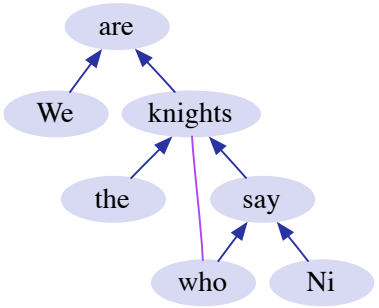
- Outputs
  - Word sentence aligned parallel data
  - Metadata about documents
  - Incremental updates

Annotation

- GFL Parser
- GFL Annotation Web Interface







- Represents unlabeled dependencies
- Special handling for:
  - multiword expressions
  - coordination
  - anaphora

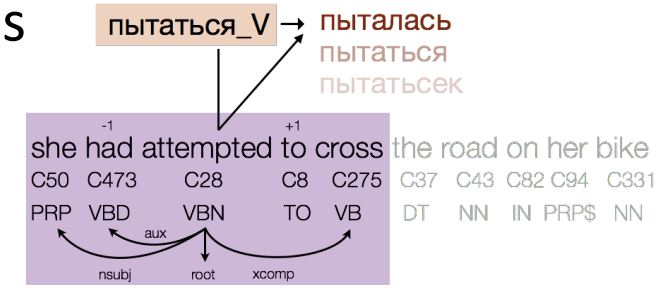
- Allows underspecification
- Graph fragment language for easy annotation

Statistical Modeling Toolkits

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

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

- Translation Modeling
  - fast_align - fast word alignment
  - spectral-scfg - fast grammar refinement for MT
  - morphogen - generate synthetic phrases



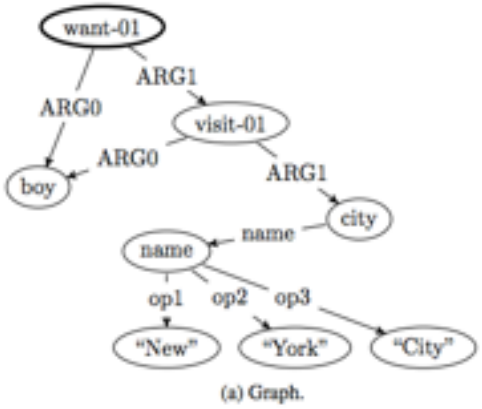
- Forthcoming
  - unsupervised CRF autoencoder suite
  - decipherment tools

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



(a) Graph.

  - yMetaphor - detect figurative language

Where?

http://www.linguisticcore.info