Basic Language Resources

Chris Cieri
Mike Maxwell
Stephanie Strassel
Low Density Languages Project

- 100k words monolingual text
- 100k words bilingual text
- 100k words text annotated for named entities
- 10k word bilingual lexicon
- Morphological parser/ stemmer
- Encoding converters
- Languages:
  Bengali, Panjabi, Tamil, Tigrinya, Uzbek, Tagalog
REFLEX Project

- Research on English and Foreign Language EXploitation
  - *Proposal stage only!*
  - Seven languages per year
  - 250k monolingual text
  - 250k bilingual text
    (75k English → target language)
  - Encoding converters
  - Sentence segmenter
  - Word segmenter (where required)
  - 10k Bilingual Lexicon
  - POS tagset and tagger
    (and for some languages, 5k word annotated text)
  - Morphological analyzer
    (and for some languages, 5k word annotated text)
  - Named entity tagger
  - 100k text annotated for named entities
Language Survey

• Languages with > 1M speakers
• Sociolinguistic status
  – Written status
  – News media
• Basic linguistic typology
• Electronic resources
  – Web sites
  – Lexicons
  – Other tools