Selection Criteria for Low Resource Language Programs

Christopher Cieri, Mike Maxwell, Stephanie Strassel, Jennifer Trace

**Background**
- Despite increased work on resource-poor language differences, in:
  - terminology
  - available information
  - goals

yield obscurity in language selection criteria

- Program goals range from:
  - LORELEI: facilitate situational awareness in the event of a disaster
  - METNET: create missing technologies and transfer languages facing digital extinction
  - NSF DEL: “document living endangered languages and their associated cultural and scientific information before they disappear”

- Program Effects
  - time & funding commitments
  - create critical language resources
  - enable human language technologies
  - increase native speaker information access

The potential impact – on research and daily life – of resource development efforts make language selection criteria a worthy topic.

**Goal:** begin dialog on how community decides which languages to study, survey selection criteria used by low resource language research and available

**Programs**
- DARPA TIDES: translilingual IR/IE & summarization in 3 languages + 1 surprise
- REFLEX LCTL: translational technologies, language packs in 20 languages
- NIST LRE: 1996-present evaluation campaign, language variety, confusability, not specifically low resource
- IARPA Babel: escape English bias in speech recognition
- DARPA LORELEI: information awareness for disaster events in low resource languages

**Terms**
- endangered: risk of losing native speakers
- critical: undesirable supply/demand ratio
- low density: few online resources (under-resourced, low resource)
- less commonly taught: e.g. specific market surprise: within common task program
- low-affluence: defined via GLP

**Selection Criteria: Resource**
- # resources
- too few mires technology development
- too many not representative
- specific resource types
- monolingual & parallel text, speech
- dictionaries, gazetteers, grammars
- human resources
- previously, local speaker population
- in-country partners
- in-country infrastructure
- elaborated
  - standard digital encoding, news & parallel text, translation dictionaries, tokenizers, segmenters, tags, morph analyzers
- different weightings of the above

**Implementation Challenges**
- different notions of ‘language’
- different language names
- difficulty collecting data on
  - demographics
  - linguistic features
  - available resources
- demographics change over time
- # Syrian Arabic speakers in Europe
- resource availability, change over time
- Quechuan in the LCTL era versus today
- language attitudes
- suppression, language death

Select References