Under-resourced languages: A personal view from industry

Martin Jansche
Google Research & Machine Intelligence

The views and opinions presented here are my own and do not necessarily reflect those of my employer.

Past activities

- GOOG-411 in India (2009-10)
- Google Voice Search (2008-2013): ASR in 50+ languages
- Early focus on global reach: Afrikaans and Zulu in 2010; Indonesian and Malaysian in 2011
- Data-mining: Web-based pronunciation mining (2008 JHU Workshop)
- In-house crowd-sourced ASR data collection driven by Linne Ha (2010-)
- In-house tool development and operations for data annotation (Hughes et al., Interspeech 2010; Ainsley et al., Interspeech 2011)
- Crowd-sourced TTS data collection (2015-)

Recurring obstacles

- Ongoing mindset challenge: Under-resourced does not mean endangered.
- Ongoing mindset challenge: Building a ASR/TTS/MT in a language from scratch is completely different from a standard i18n/l10n effort.
- Ongoing need for native speaker expertise, linguistics expertise, and software development expertise. Different models tried for bringing all skills to a team; no clear winning formula.
- Ongoing need for basic resources: text corpora, tokenizers and segmenters, morphological analyzers, pronunciation lexicons and phonological descriptions.

Opportunities for collaboration

Speech technology is increasingly becoming commoditized. Basic ASR/TTS technology is part of a larger product offering, not a product in itself. Language parity remains important: need to reduce its cost. Areas for collaboration:

- Active open-source development of basic resources
- Standardize the format in which resources are delivered
- Integrate with open-source ecosystem (Kaldi, Festival, NLTK, Moses, …)
- Bake-offs / competitions / workshops aimed at resource bootstrapping