WGs 3&4: endangered languages for computational linguistics for ELs

WG3: before lunch

- CL can target various stages of the documentation process
- An important research theme: cross-lingual transfer, language similarity
  - adapt tools/methods to new languages
  - not just linguistically related but also typologically similar

Low-hanging fruit

- create mailing list for CL people working on ELs
- get people on corpora mailing list
- create a website/Google doc for sharing info about existing tools, research projects, etc.
- outreach to ACL: “low-resource languages and endangered languages”

Data collection/sharing

- data from field linguists: issues around community privacy/sensitivity, issues around linguists (data not well-organized, etc.)

- data from publicly available sources: web, FLEX database system (archiving capacity), xling for writing papers

- requirements for data sharing, and with metadata? (journals, etc.)

Clean data

- Can CL offer help with data clean up?
- Need for field data to have well-defined structure, consistent annotations, using standard terminology
- there *are* existing tools for doing data clean-up
- links to follow

Biggest barriers

- Access to data
- Funding: paths to funding opportunities differ depending on situation (i.e. students vs. faculty/PIs vs. NGOs vs. speaker communities)
- Finding scientific contributions to be made while also supporting language documentation
- Collaboration: many types of experts needed
- Evaluation

Crowd-sourcing?

- Means for involving speaker communities
- Perhaps could be helpful for data collection
- Perhaps also for annotation: typing and correcting IGT, community dictionaries, etc.

Other areas of CS

- Game with a purpose
- Work in HCI
- Web GUI development for data collection and annotation