ComputEL

Working Group 1: Tool Usability and Sustainability

Participants: Laura Welcher, Damir Cavar (co-chairs), Martin Benjamin, Jordan Lachler, Jennifer Serventi, Worthy Martin, Lane Schwartz, Daniel Fox, Steven Tratz

Needs / Barriers: Support for Tool Creation and Ongoing Development

- Institutional support is not a long-term guarantee, tool creators may not have the resources in the future to be tool sustainers.

- It isn't clear that the go-to place -- institutional repositories -- are the best places to store software/code/data. Increasingly, costs for keeping tools alive can involve off-site storage (many institutions are outsourcing their IT) and software/data projects don't have a long-term budget for dissemination/archiving.

- The "there's no-market" myth?: Collectively, the total number of endangered language /low-resource language speakers constitute a significant market, and can provide commercial value for tool development.

- But this means that repurposing tools across languages is important. It isn't clear how difficult / easy this would be to achieve.

Needs / Barriers: Open Endangered Language Resources

- There are many benefits from the open source/data revolution in computer science and computational linguistics -- a similar revolution is needed in documentary linguistics.

- Cultural / public relations issue: expectation that language resources created by endangered language documentation projects are restricted, and there are often many hurdles to access archived resources. Computational Linguists will go to where the data is and is most readily available.

- Need a showcase for the life cycle showing the creation of open corpora to tool development that can bring benefit to endangered language speech communities.

- Need a way for communities to engage with the whole process of data and tool creation to curation -- there are major broader impacts / capacity building opportunities here.

How Can We Help Each Other?

- Documentary Linguists:
  - Create open language resources, make them available and discoverable.
  - Tell computational linguists about the tools / modules that exist for their workflows, so that plugins or modifications could happen rather than developing a new tool from scratch (e.g. all of the tools that have been developed by SIL)
- Computational Linguists:
  - Provide tools and instructions for communities to repurpose tools for their particular languages -- “here are the tools, here’s how you can use them, here is what you can get if you do.
  - Provide information about tools such as: which tools are used for what kind of language? (in terms of structure), what size corpus is needed?
  - Provide learning materials / curricula for how you can learn to use and enhance tools.

Low-hanging Fruit - Missing Tools and Communities of Practice

- Localization for open-source software. Community members who have an interest in language revitalization could work on this (might require substantial vocabulary creation though).

- Data conversion tools (e.g. from ELAN XML to...) and standards for interoperability

- Tools that speed up any aspect of the documentation workflow - creation of primary resources, the transcription bottleneck.

- Tools to archive data throughout the data life cycle. An “archive this” button. Versioning to support “archive early, archive often”.

- Need for active user communities of developers to identify tools and platforms where more work needs to be done.

- Software registries / repositories -- bringing endangered language software developers into the domain of software preservation.

Low-hanging Fruit -- Fundable Projects

- Case studies for sustainability -- identify other projects out there that have sustainable models for tool development and maintenance. Identify special needs that tool development / maintenance has for low resource, endangered languages. [See N. Maron (2014) “A Guide to the Best Revenue Models and Funding Sources for your Digital Resources”]

- Showcase the need for and benefits of creating open language resources. Show the life cycle showing the creation of open corpora to tool development that can bring benefit to endangered language speech communities.

- Need a way for communities to engage with the whole process. Major broader impacts / capacity building opportunities: community-based learning of computational linguistics – getting skills into the hands of the community, parallel to what has been done with bringing documentation skills into communities.