Day 2/Working Group 2
Annotation and transcription practices
Our original charge

- Compare the annotation systems and content foci in order to determine the consequences, if any, of different choices of annotation tools.
- How difficult would it be to switch from these systems to one of the systems used for automatic speech recognition?
- Early conclusion: When considering format alone, the problems are not that great
Annotation content

• Much of the workshop has focused on transcription—a key kind of annotation

• But, there are other kinds as well

  • Language identification
  • Speaker identification/diarization
  • Silence detection
  • Speaker emotional state
  • ...

...
Annotation possibilities

• Hard right now: Automatic speech recognition for an arbitrary language

• Easier right now
  • Speaker turns
  • Language identification
  • Speaker identification
  • Automatic transcription of majority metalanguage of a recording
  • Keyword spotting
Annotation services

- Would AARDVARC services using mature technologies encourage people to deposit?
- Automatic speaker identification could help improve metadata for a recording
- Automated transcription of English could help with transcription of elicitation
- Silence detection could enable a “hide silence” feature in an annotation tool
- Implementable services today may give us data we need for better services tomorrow
Other resources

- Many advanced transcription and text processing techniques rely on a lexicon
- Depending on the language, a model of morphology may also be needed
- AARDVARC, building on LEGO, could offer a lexicon quality control service
- Additional services?
  - Morphological model creation
  - Outputs for different audiences (e.g., community members)
  - Transcoding systems
Feedback loop

• As useful services cause people to add data, the dataset becomes more interesting for the development of new services

• We can start with turn-taking, but this can help lead to better automated transcription

• This model does not produce cutting edge research initially, but it could build a community of depositors

• This alone may not be sufficient for funding, but it could be a useful component
AARDVARC model

- This approach to annotation implies a certain general model for AARDVARC
  - A repository that would hold data for those without access to another repository
  - A set of AARDVARC-compliant services for working with uploaded data
  - A mechanism for the registration of new services
  - A “front-end” for data processing pipelines
The repository

- One possible repository: The LDC
- It will digitize tapes! It has more capacity than people are making use of.
- The metadata barriers to deposit materials should be as minimal as possible
- AARDVARC services could try to automatically enrich the metadata
The services

• Many key tools already exist, but people don’t know how to find them

• CLARIN provides a model, but, as a catalog, its utility for the data provider is limited

• Wherever the tools may exist, they should be accessible from one place

• And, it should be possible to join them into a processing pipeline in a standardized way
New services

• As the dataset becomes richer, it should be possible to add new services

• There should be a standardized way to fit these into pipelines

• There should be a standardized way to register them as available

• Even language-specific services should be available since they may be adaptable for other languages
An existing model: UIMA
A UIMA pipeline
Summary

• There’s more to annotation than transcription
• Some useful services could be built quickly
• The more data that comes in, the more attractive the repository is for developers
• AARDVARC can be built around
  • A repository
  • Services
  • A system for managing pipelines of services