Filtering and annotating web speech data

WAC 8
Lancaster, UK
July 22, 2013

David Lutz
Parry Cadwallader
Mats Rooth

Cornell University
Research with Web Speech Data

1. Locate
2. Collect
3. Filter
4. Annotate
5. Analyze
Web Speech Data

1. Locate
2. Collect
3. Filter
4. Annotate
5. Analyze

Howell and Rooth 2009;
Howell 2012
Web Speech Data

1. Locate
2. Collect
3. Filter
4. Annotate
5. Analyze

Today: ezra
Locating and collecting tokens

- Content providers offer “media search”
- Many results, often with transcripts
- Results come from text search of ASR-generated transcripts
Locating and Collecting tokens

- Howell and Rooth: tools to search and download results
- Yielded hundreds of hits for target phrases
- Many of these hits were false positives
- Those that weren’t needed annotation
Filtering and Annotation

1. Separate the true tokens from the false positives / duplicates
2. Mark the token location and reasonable clip boundaries
3. Add or correct the clip transcript
4. Target-specific annotation tasks

Much time was spent repeating the same mindless tasks
Work is organized around **targets** and **features**.

**targets** are words or phrases of interest to the researcher.
Work is organized around targets and features.

features are annotation tasks for a particular target.
Features

- Features are annotation tasks
- Associated with targets
- Annotators complete these tasks for each hit
- Feature values are stored with each hit
ezra: filtering and annotation

any players 6964

Features
Hide All

Det N

prosody

since the Bruins as you point out or or as you believe remain committed to Claude Julien uh do you envis do you see there being any players on the horizon anyone available via trade who could help the team

Keyboard shortcuts

Notes:
Workflow

1. Define a target and import data
2. Define features and associate them with target(s)
3. Filter and annotate hits
4. Export for analysis
Two user classes

- Supervisors
  - Create targets and features
  - Import/export data
  - Monitor user activity
  - Supervise annotation

- Annotators
  - More limited privileges
  - Focus on filtering and annotation
Web application

- Accessible from anywhere
  - Users need a modern browser and internet connection
  - Team members can be remote

- Standard technology
  - Ruby on Rails
  - HTML5 / Javascript
  - SQLite
  - Works in modern browsers
Ezra: benefits

• Web application
• Built on standard technology
• Two user classes
• Multiple targets/projects

• Flexible feature definition
• Simple interface
• Public site
• Efficiency

https://github.com/del82/ezra
Ezra: future work

- More complete user auditing
- Integrate locating and collecting tokens: plugins
- Improve administrator interface
- Automatic duplicate detection
- Partnerships with content creators
- Crowdsourcing?

https://github.com/del82/ezra
Thanks

Neil Ashton
Jonathan Howell
Michael Wagner
Ross Kettleson
Michael Schramm
Lauren Garfinkle

NSF 1035151 RAPID: Harvesting Speech Datasets for Linguistic Research on the Web (Digging into Data Challenge)