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

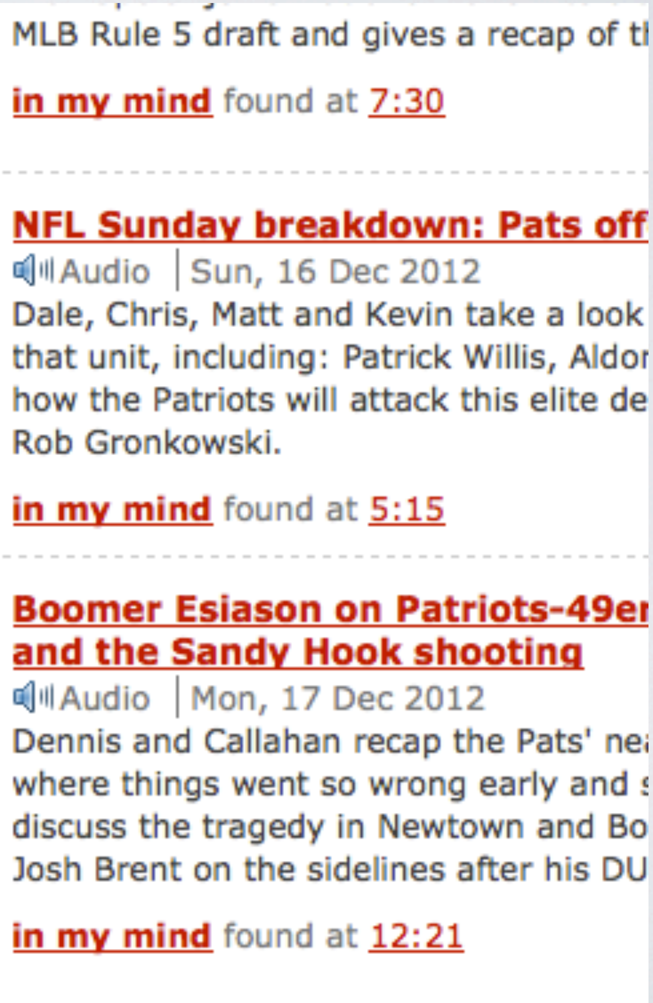
4. Annotate

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



MLB Rule 5 draft and gives a recap of t

in my mind found at 7:30

NFL Sunday breakdown: Pats off

🔊Audio | Sun, 16 Dec 2012

Dale, Chris, Matt and Kevin take a look
that unit, including: Patrick Willis, Aldo
how the Patriots will attack this elite de
Rob Gronkowski.

in my mind found at 5:15

Boomer Esiason on Patriots-49e
and the Sandy Hook shooting

🔊Audio | Mon, 17 Dec 2012

Dennis and Callahan recap the Pats' ne
where things went so wrong early and
discuss the tragedy in Newtown and Bo
Josh Brent on the sidelines after his DU

in my mind found at 12:21

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

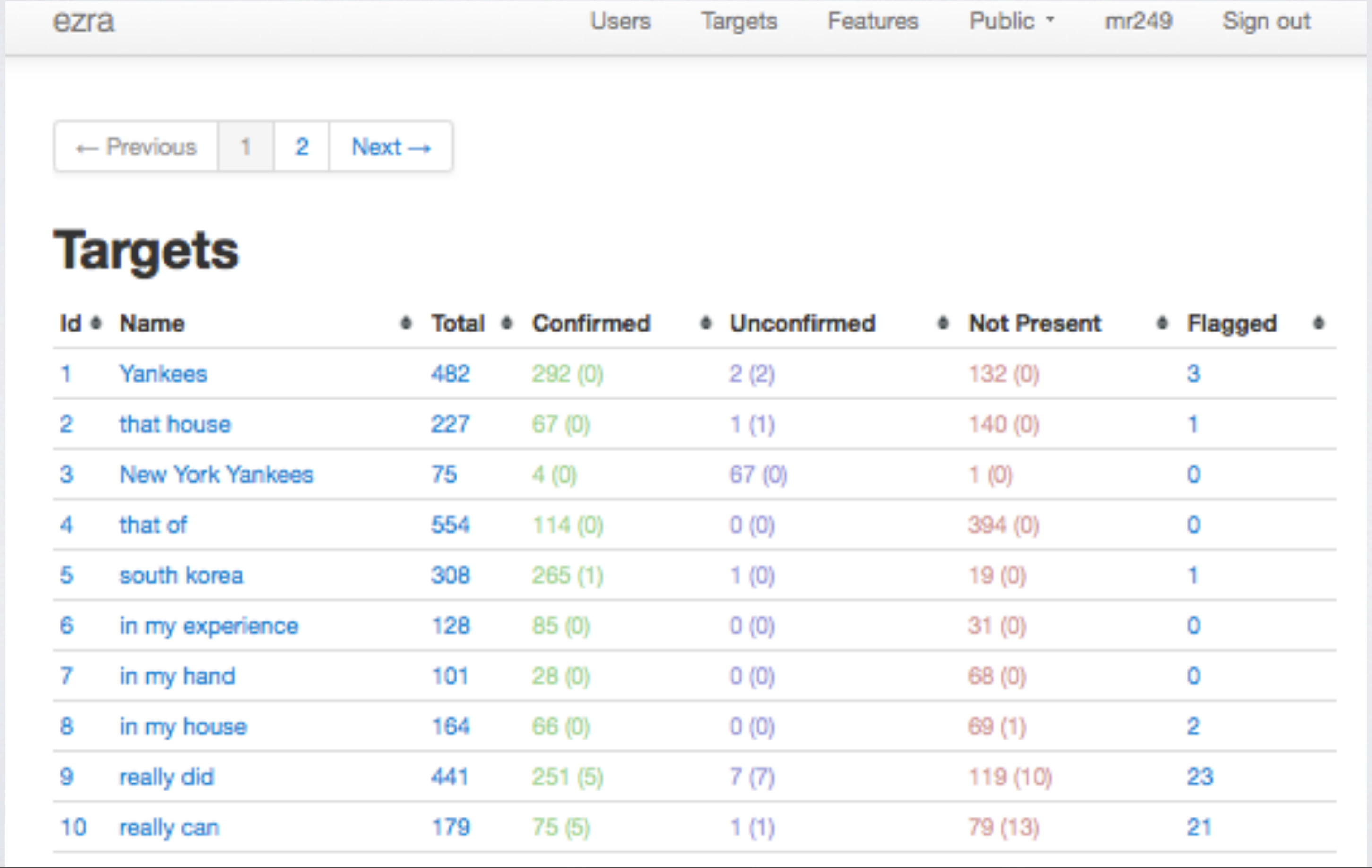
1. Separate the true tokens from the false positives / duplicates
2. Mark the token location and reasonable clip boundaries
3. Add or correct the clip transcript
4. Target-specific annotation tasks

Much time was spent repeating the same mindless tasks

ezra

Work is organized around **targets** and **features**.

targets are words or phrases of interest to the researcher.



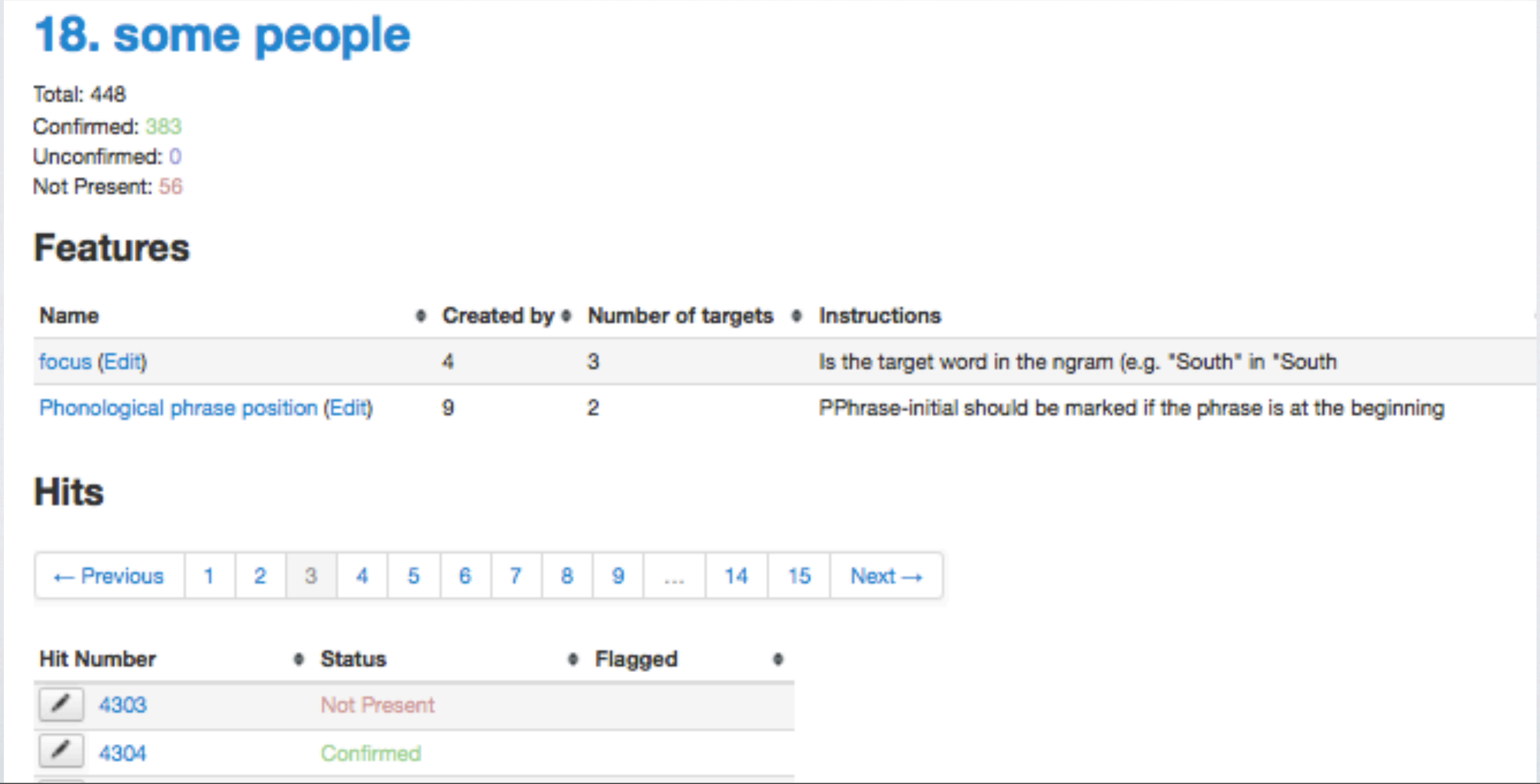
Targets

Id	Name	Total	Confirmed	Unconfirmed	Not Present	Flagged
1	Yankees	482	292 (0)	2 (2)	132 (0)	3
2	that house	227	67 (0)	1 (1)	140 (0)	1
3	New York Yankees	75	4 (0)	67 (0)	1 (0)	0
4	that of	554	114 (0)	0 (0)	394 (0)	0
5	south korea	308	265 (1)	1 (0)	19 (0)	1
6	in my experience	128	85 (0)	0 (0)	31 (0)	0
7	in my hand	101	28 (0)	0 (0)	68 (0)	0
8	in my house	164	66 (0)	0 (0)	69 (1)	2
9	really did	441	251 (5)	7 (7)	119 (10)	23
10	really can	179	75 (5)	1 (1)	79 (13)	21

ezra

Work is organized around **targets** and **features**.

features are annotation tasks for a particular target.



18. some people

Total: 448
Confirmed: 383
Unconfirmed: 0
Not Present: 56

Features

Name	Created by	Number of targets	Instructions
focus (Edit)	4	3	Is the target word in the ngram (e.g. "South" in "South
Phonological phrase position (Edit)	9	2	PPhrase-initial should be marked if the phrase is at the beginning

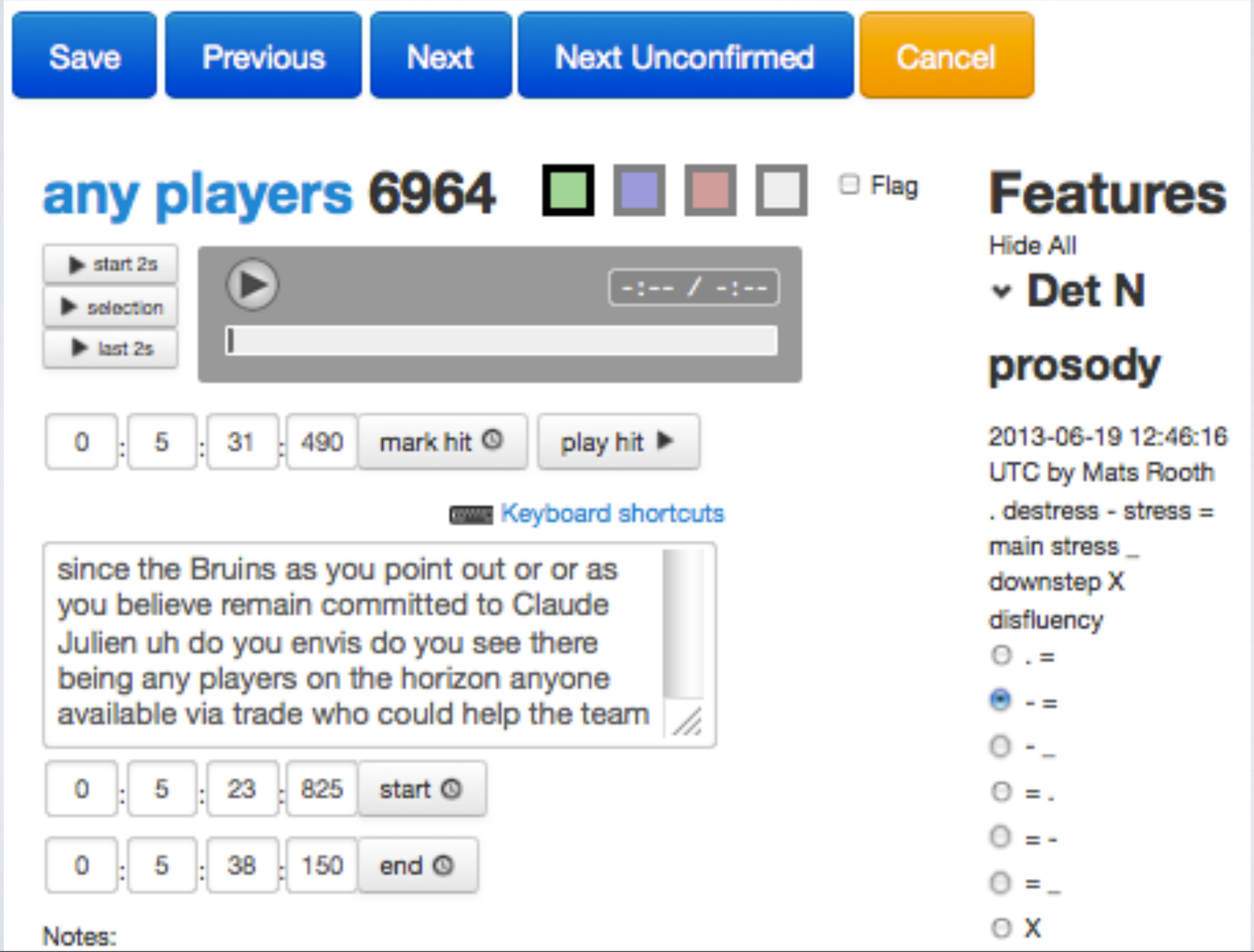
Hits

Hit Number	Status	Flagged
4303	Not Present	
4304	Confirmed	

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 🟩 🟪 🟥 ⬜ □ Flag

0 : 5 : 31 : 490 mark hit

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

0 : 5 : 23 : 825 start

0 : 5 : 38 : 150 end

Notes:

Features

Hide All

⌄ Det N

prosody

2013-06-19 12:46:16 UTC by Mats Rooth

. destress - stress = main stress _ downstep X disfluency

○ . =
● - =
○ - _
○ = .
○ = -
○ = _
○ X

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)