

---

*Taking aim at the transcription bottleneck:*  
**Integrating speech technology into language  
documentation and conservation**

---

Christopher Cox (*Carleton University*)

Gilles Boulianne (*Centre de recherche informatique de Montréal*)

Jahangir Alam (*Centre de recherche informatique de Montréal*)



**LD&C**  
2019

# Acknowledgments

*We gratefully acknowledge the intellectual contributions of colleagues, collaborators, and friends in many communities, including speakers and learners of Tsuut'ina and Plautdietsch, as well as the financial support of the National Research Council of Canada, CANARIE, and the Trudeau Foundation for aspects of this work.*



**Carleton**  
UNIVERSITY



**CRIM**



**canarie**



F O N D A T I O N  
**TRUDEAU**  
F O U N D A T I O N

# Setting the scene

- Three related observations:
  1. Transcription (*and translation*) is a **central activity** in language documentation and conservation (LDC)—but rarely a **central focus** of documentary linguistic discussions<sup>1</sup>

---

1. Jung & Himmelmann (2011: 201)

“

[T]he **transcription and further annotation** of recordings (...) constitute the **major workload** in a documentation project.

”

*Himmelmann (2008: 347; emphasis added)*

“

It is only a minor exaggeration to say that language documentation is all about transcription.

”

*Himmelman (2018: 38; emphasis added)*

“

Transcribing narrative and conversational speech is a **core activity** of all linguistic fieldwork, though **one of the less attractive ones**. (...) Nevertheless, it is without doubt one of the most important tasks to be carried out in the field requiring close cooperation between speaker(s) and researcher(s).

”

*Jung & Himmelmann (2011: 201; emphasis added)*

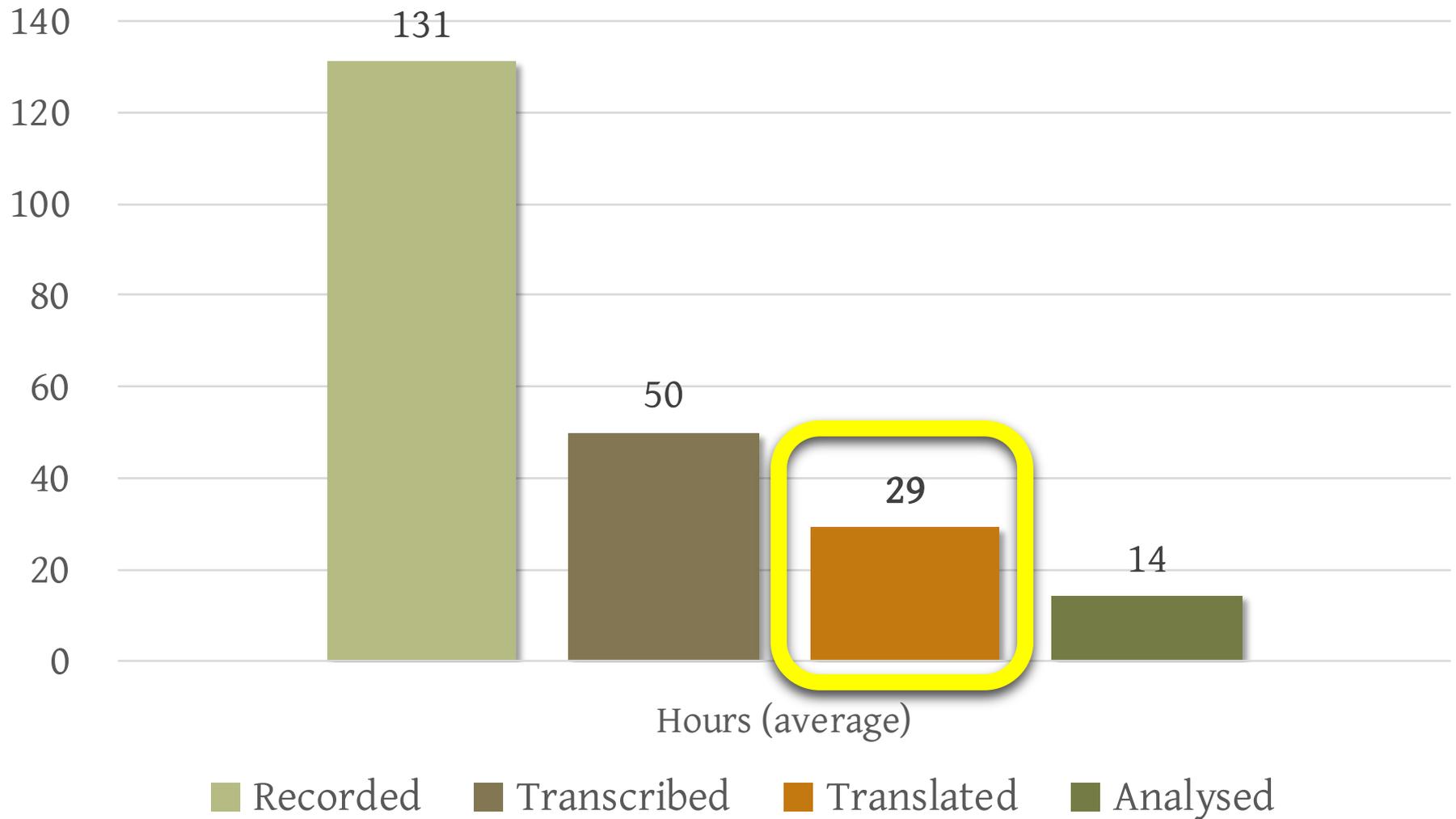
# Setting the scene

- Three related observations:
  1. Transcription (*and translation*) is a **central activity** in language documentation and conservation—but rarely a **central focus** of documentary linguistic discussions<sup>1</sup>
  2. Our **ability to record** (*and archive*) language materials **outstrips** our ability to **make their contents accessible** through transcription<sup>2</sup>

---

1. Jung & Himmelmann (2011: 201)    2. Himmelmann (2006)

# DOBES Projects (2000-2009)



*(after Wittenburg (2009: slide 34), cited in Austin 2010)*

# The ubiquitous backlog

- *Elephant in the room:* Most language documentation and conservation initiatives that involve recording end up with a **backlog of unannotated, 'raw' recordings**
  - Not at all unusual, but generally not discussed (*at least not in public*)
  - Issue for both documentation and revitalization-focused initiatives

## *Example: Consultation sessions*

- Language meetings with speakers of Tsuut'ina (Na-Dene; ISO 639-3: *srs*)
  - Sessions **multilingual** (*Tsuut'ina, English*) and **multi-speaker** (*2-3 people in meeting*)
  - Meetings typically recorded (*cf. Jung & Himmelmann 2011*)—**200+ hours** of audio
  - Recordings valuable, but contents difficult to access due to extent (*not feasible to manually segment, even only Tsuut'ina-language portions; time investment for oral annotation prohibitive*)

# Setting the scene

- Three related observations:
  1. Transcription (*and translation*) is a **central activity** in language documentation and conservation (LDC)—but rarely a **central focus** of documentary linguistic discussions<sup>1</sup>
  2. Our **ability to record** (*and archive*) language materials **outstrips** our ability to **make their contents accessible** through transcription<sup>2</sup>
  3. Addressing the resulting accessibility issues is a **standing challenge** for current work in LDC<sup>3</sup>

---

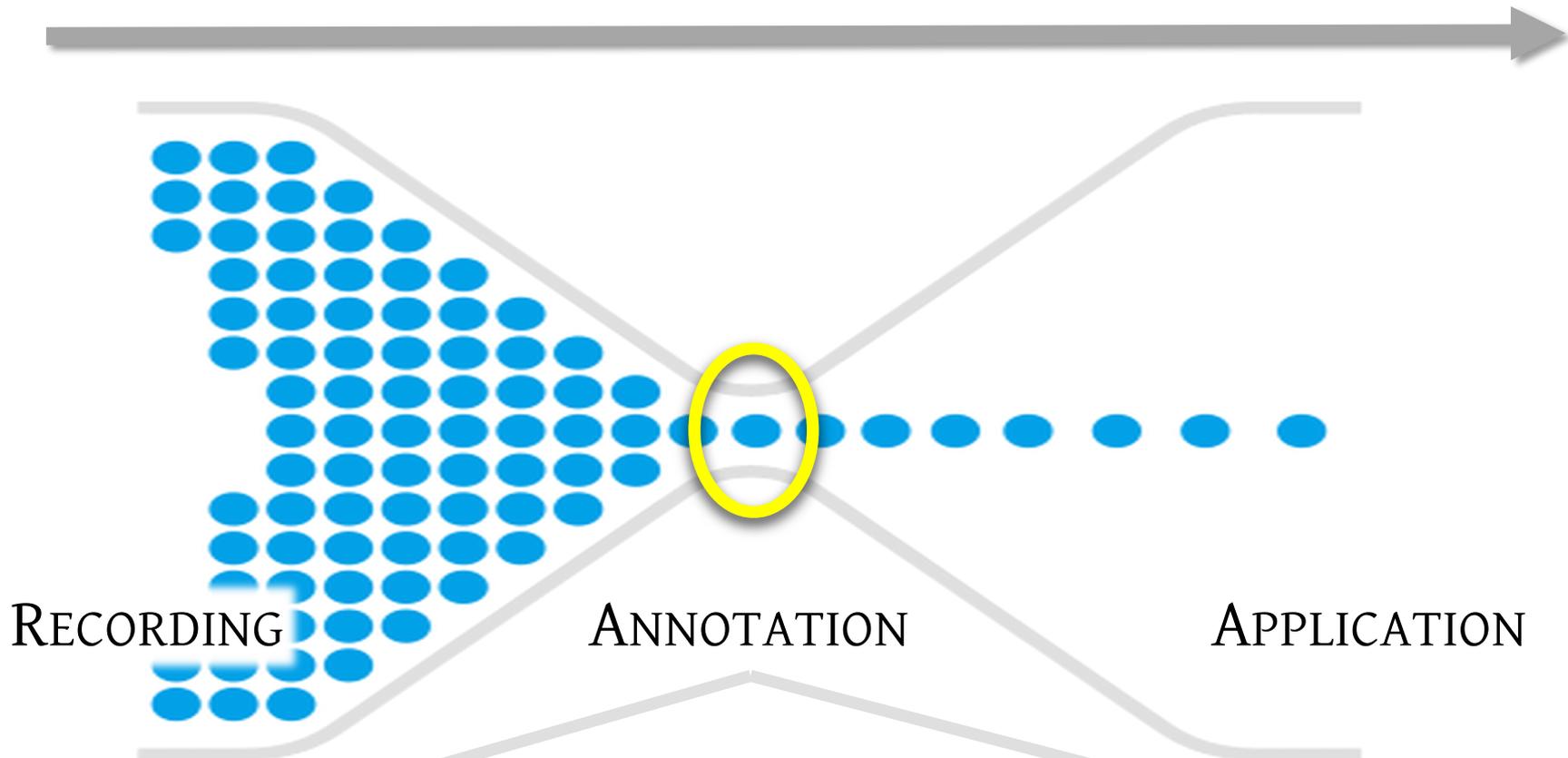
1. Jung & Himmelmann (2011: 201)   2. Himmelmann (2006)   3. Thieberger (2016)

# Annotation and documentary linguistics

- Difficulties in providing ‘baseline’ bilingual annotation as possibly informing the direction that some areas of LDC are currently taking:
  - **Movement away from projects involving extensive recording** and relatively little baseline annotation (*e.g., 100–200 hours recorded, 10–20% bilingually annotated*) and **towards smaller projects** with much more annotation (*e.g., 10–20 hours recorded, 90–100% bilingually annotated*)<sup>1</sup>

---

1. Austin (2017)



Segmentation  
*(into utterances)*

'Baseline'  
annotation

Further  
annotation

# Current approaches

- Thankfully, several **emerging methods and technologies** to addressing different aspects of this bottleneck:
  - *Oral annotation: **BOLD** (Reiman 2009, 2010, Boerger 2011, Boerger et al. 2019), **Aikuma** (Bird et al. 2014, Adda et al. 2016, Gauthier et al. 2016), inter alia*
  - *Written annotation: **Automatic phoneme recognition** (e.g., *Persephone*: Adams et al. 2018, Michaud et al. 2018, Cox et al. 2018), **automatic speech recognition** (e.g., Foley et al. 2018, Jimerson & Prud'hommeaux 2018, inter alia)*

# Looking outside of LDC

- Many other areas of research (*and digital life, more broadly*) face similar challenges in addressing this ‘data deluge’:
  - Similar challenges in oral history research, corpus linguistics, media production, etc.
  - Increasingly addressed with methods from **computational linguistics** and **natural language processing** (*e.g., automatic captioning of YouTube videos in several languages*)

# Looking outside of LDC

- Until recently, these methods have largely seemed **out of reach** for smaller or lesser-studied languages, with CL/NLP research focused on larger languages with extensive digital resources<sup>1</sup>
- Thankfully, that has begun to change:
  - *People*: Increasing interaction between computational linguistics and LDC (e.g., ComputEL)
  - *Tools*: Integration of **web services** (*text-focused*) and **recognizers** (*A/V-focused*) into common documentary tools like ELAN

---

1. Bird (2009)

# CRIM-Carleton collaboration

- **CRIM:** Expertise in development of speech technologies (*e.g., state-of-the-art automatic speech recognition for Canadian French/English*)
- Existing web-based platform, **VESTA**, into which speech technologies had been integrated to support social science and education research (2014-)





## Speaker segmentation

Partition an audio frame into segments according to the identity of the person speaking. Vesta provides facilities to determine the person is speaking and the gender.



## Speech-to-text

Transcribe the speech of an audio frame. A domain specific vocabulary can be provided to improve the results.

# CRIM-Carleton collaboration

- **CRIM:** Expertise in development of speech technologies (*e.g., state-of-the-art automatic speech recognition for Canadian French/English*)
  - Methods robust, implemented as web services that could be called from anywhere—but not previously applied to lesser-studied languages
  - **Q:** Could these same functions be **integrated into common documentary linguistic workflows?**

# Introducing VESTA-ELAN

- *Idea:* Integrate VESTA services directly into ELAN for easier use in documentation projects:
  1. **Automatic segmentation:** Identify speech vs. non-speech sections of recordings (*language-independent task*)
  2. **Speaker diarization:** Attribute speech sections to different speakers (*language-independent task*)
  3. **Content language identification:** Recognize which segments are primarily English and which aren't (*language-dependent task; work in progress*)
  4. **Automatic speech recognition:** Transcribe any speech in English or French (*language-dependent task*)

# Example 1: VESTA diarization

The screenshot displays the ELAN 5.4 software interface, specifically the 'Recognizers' tab. The window title is 'ELAN 5.4 - 2010-07-22-srs-DR-Wetaskiwin-Korg-3-EXAMPLE.eaf'. The menu bar includes 'Grid', 'Text', 'Subtitles', 'Lexicon', 'Comments', 'Recognizers', 'Metadata', and 'Controls'. The 'Recognizer' dropdown is set to 'AAM-LR Phone level audio segmentation'. The 'Parameters' section is expanded, showing 'Settings' with the 'Base URL of AAM-LR service' set to 'http://lux17.mpi.nl/aamlr/'. The 'Input' section shows the audio file '2010-07-22-srs-DR-Wetaskiwin-Korg-3-ELAN\_41280\_68520.wav'. The 'Output' section shows the tier 'Tier holding the phone level segmentation'. The 'Progress' section has a progress bar and buttons for 'Start', 'Report...', and 'Create Tier(s)...'. The playback control bar shows 'Elapsed time: 00:00' and 'Time since last update: 00:00'. The timeline view shows the audio waveform and diarization segments labeled 'DR' and 'CDC'.

# Example 2: VESTA speech recognition

ELAN 5.4 - 2010-07-22-srs-DR-Wetaskiwin-Korg-1-EXAMPLE.eaf

Grid Text Subtitles Lexicon Comments **Recognizers** Metadata Controls

Recognizer: AAM-LR Phone level audio segmentation

Parameters

Settings

Base URL of AAM-LR service  
http://lux17.mpi.nl/aamlr/

Input

[audio]: Input audio file  
2010-07-22-srs-DR-Wetaskiwin-Korg-1-ELAN\_2393510\_2417300.wav

Progress

Elapsed time : 00:00 Time since last update : 00:00

Start Report... Create Tier(s)...

00:00:00.466 Selection: 00:00:00.000 - 00:00:00.000 0

2010-07-... 00.000 00:00:01.000 00:00:02.000 00:00:03.000 00:00:04.000 00:00:05.000 00:00:06.000

DR [0]

# Example 3: VESTA + Other recognizers

The screenshot displays the ELAN 5.4 software interface, specifically the 'Recognizers' tab. The window title is 'ELAN 5.4 - OS-srs-BRS-040-2015-07-18\_0900.eaf'. The 'Recognizer' dropdown is set to 'AAM-LR Phone level audio segmentation'. The 'Parameters' section is expanded, showing three sub-sections: 'Settings' with a 'Base URL of AAM-LR service' field containing 'http://lux17.mpi.nl/aamlr/'; 'Input' with an '[audio]: Input audio file' field containing 'OS-srs-BRS-040-2015-07-18\_0900-ELAN.wav'; and 'Output' with an '[xml tier]: Tier holding the phone level segmentation' field. Below the parameters is a 'Progress' section with a progress bar, 'Elapsed time: 00:00', 'Time since last update: 00:00', and buttons for 'Start', 'Report...', and 'Create Tier(s)...'. The bottom section of the interface features a playback control bar with a time display of '00:00:02.820' and a selection range of '00:00:00.000 - 00:00:00.000 0'. Below the controls is an audio waveform visualization with a time axis from 00:00:05.000 to 00:00:08.500. A red shaded area at the bottom of the waveform is labeled 'BRS [0]'.

# Conclusions

- VESTA-ELAN services target a particular range of issues in the current transcription bottleneck, aiming to **make written annotation more feasible**
  - Sets the stage for further **automatic** and **semi-automatic annotation** to be applied (*e.g., first-pass phonemic transcription using Persephone; cf. Adams et al. 2018, Cox et al. 2018*)

# Conclusions

- The VESTA-ELAN recognizers will be **made generally available for public use** soon
  - Aim to be a useful addition to the LDC toolkit—one that facilitates both ‘traditional’ transcription/translation and automatic annotation techniques
  - Integration with other, similar annotation services currently under development may help reduce the “transcription bottleneck”—encouraging more expansive documentation projects than may currently be feasible.

*Thanks!*