Comput*EL*

The use of computational methods in the study of endangered languages

52nd Annual Meeting of the Association for Computational Linguistics

Baltimore, Maryland

26 June 2014

Comput*EL*: The need

- Three concerns of NSF's Documenting Endangered Languages program
  - Data (e.g., new recordings)
  - Infrastructure (e.g., digital archives)
  - Computational methods (and tools)
- The impact of computational methods has been less than originally expected
- The problems is not lack of interest, but disciplinary cultures

Two-part event

- Two components to Comput*EL*
  - An official ACL workshop (today)
  - A "closed" meeting (tomorrow)
- The goals
  - Learn about the state of the art (from both the EL side and the CL side)
  - Assess past work and develop a future agenda
- A diverse set of attendees

Overview of workshop day

- Paper schedule should be available—note slight change from schedule in proceedings
- Three paper sessions and a poster/demo session from 2:00–3:00
- Demos for following papers: Beale, Bender et al., Bird et al., and Dunham et al., Snoek et al., Ulinski et al.
- One poster not on program: Binyam Gebrekidan Gebre of the The Language Archive
- ACL refreshments available to all registered attendees; lunch is on your own from 12:20–2:00
- Those attending tomorrow: We are not entitled to the ACL refreshments (sorry)

Overview of meeting day

- Plenary sessions and working groups
  - WG1: Tool usability and sustainability
  - WG2: Community building
  - WG3: Computational methods for ELs
  - WG4: Contribution of ELs to CL
- Prospective orientation
- Ideal outcome: New collaborations at the intersection of ELs and CompLing
- Location: This room?

Acknowledgments

- My co-organizers Julia Hirschberg and Owen Rambow
- The workshop program committee
- The ACL meeting organizers
- The National Science Foundation
  (Award Nos. BCS-1404352 and IIS-1027289)

Comput*EL*

The use of computational methods in the study of endangered languages

52nd Annual Meeting of the Association for Computational Linguistics

Baltimore, Maryland

27 June 2014

Comput*EL*: The need

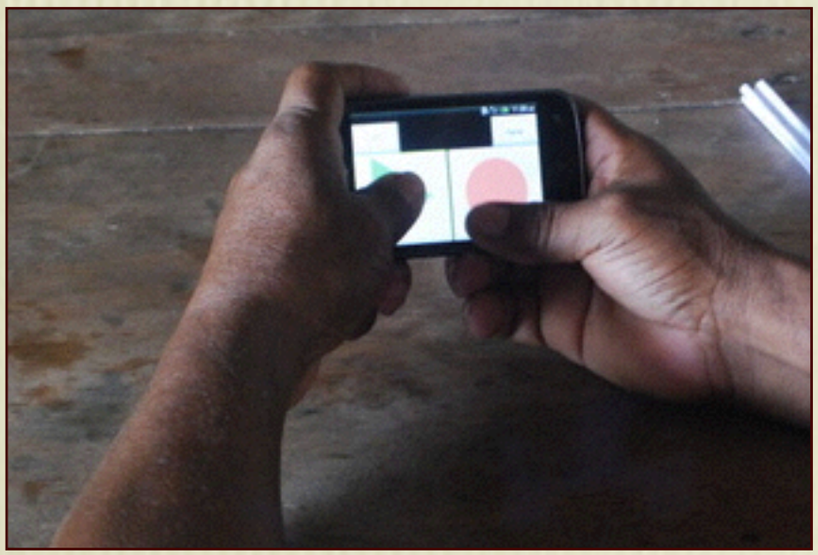
- Three concerns of NSF's Documenting Endangered Languages program
  - Data (e.g., new recordings)
  - Infrastructure (e.g., digital archives)
  - Computational methods (and tools)
- The impact of computational methods has been less than originally expected
- The problems is not lack of interest, but disciplinary cultures

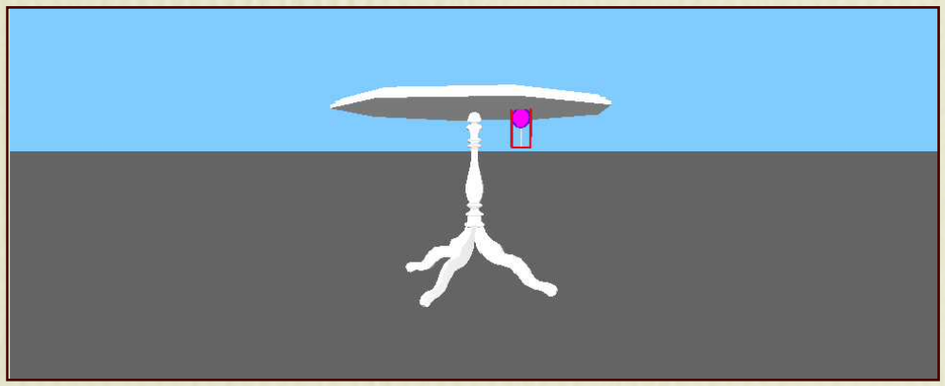
Two-part event

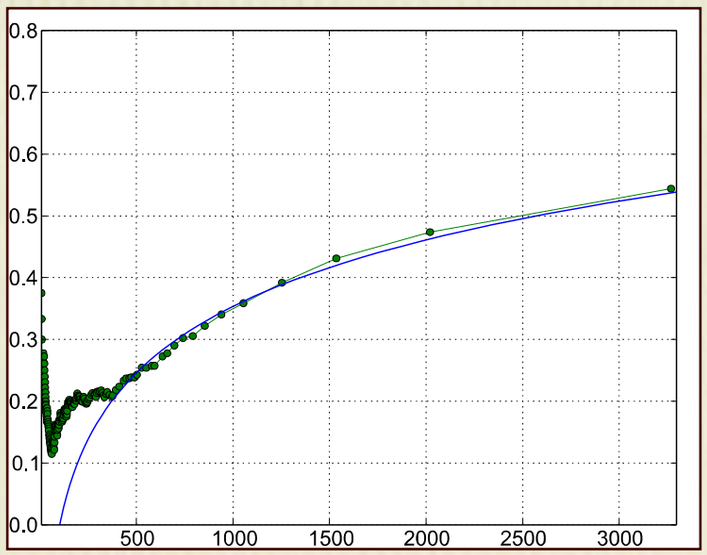
- Two components to Comput*EL*
  - An official ACL workshop (today)
  - A "closed" meeting (tomorrow)
- The goals
  - Learn about the state of the art (from both the EL side and the CL side)
  - Assess past work and develop a future agenda
- A diverse set of attendees

Overview of ACL workshop day

- Held yesterday, fourteen papers and posters
- Proceedings available at: http://acl2014.org/acl2014/W14-22/
- Papers from the EL and CL side
- Topics from tools under development, to training, to programmatic concerns







Overview of meeting day

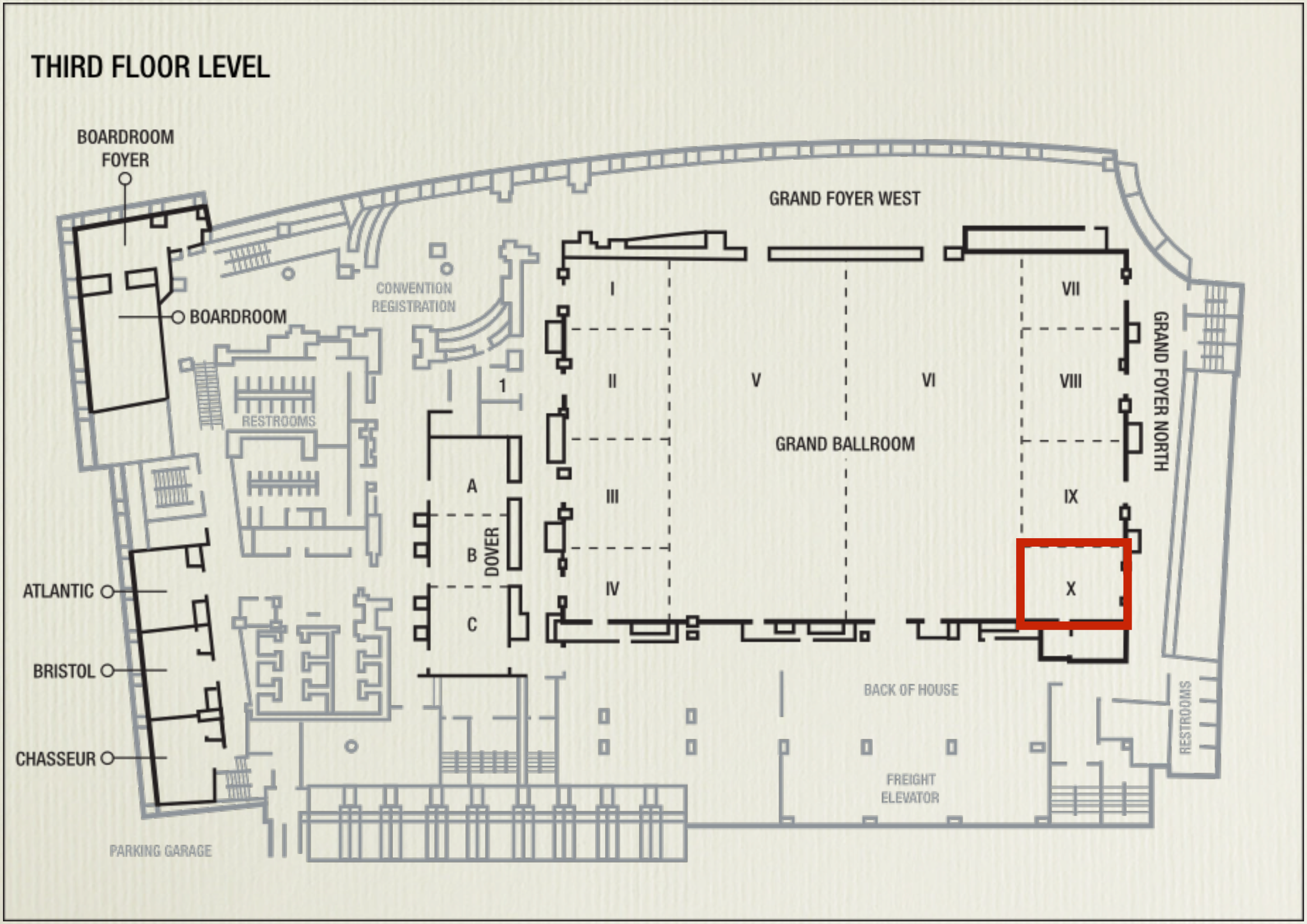
- Plenary sessions and working groups
  - WG1: Tool usability and sustainability
  - WG2: Community building
  - WG3: Computational methods for ELs
  - WG4: Contribution of ELs to CL
- Prospective orientation
- Ideal outcome: New collaborations at the intersection of ELs and CompLing

Today's logistics

- WG1: Here
- WG2: Here
- WG3: Chausseur
- WG4: Bristol

- Begin with brief introductions
- Two sessions of short presentations
- We aren't entitled to ACL refreshments—but, if allowed, you can have a $10 Starbucks gift card
- Lunch will be on your own (12:00–1:30)

Grand Ballroom X



Folder contents

- In your folders, you will find
  - The attendee list (including some who could only attend yesterday)
  - Today's schedule and description of the working groups
  - An edited version of the ACL workshop proposal for reference

Scaling up the data

- What does the future philologist need?
- What do NLP experts need today? Tomorrow?
- How can multilingual data inform EL work?
- How can standard EL practices (e.g., IGT) be exploited and/or refined for CL purposes?
- EL linguists are language/community specialists, depth of a language over breadth
- CL linguists can focus on broad solutions (e.g., many languages) over in-depth ones
- Highly diverse samples may extend to Els

Workflow model

- Diverse actors: Linguist, community member, technician, etc.
- Domain general problems: Data integration, curation, security
- Full-blown implementations become difficult, complex; seem to require complex fieldwork
  - Do we need new fieldwork models?
  - Community tools ↔ research tools
- When can a single tool do two jobs: Improve workflow and document something new

EL properties

- Lack of standardization; lack of texts
- Different sociolinguistic context may produce different grammatical patterns
- Linguistic activities more likely to have community impact (endangered, or threatened?)
- …?

Cultural barriers

- How many people have a good understanding of field work and computational linguistics?
- Without this understanding, collaboration (direct or indirect) is difficult
- Training of students is a big part of this; students are good at training each other
- EL data collectors not driven to be data sharers

Acknowledgments

- My co-organizers Julia Hirschberg and Owen Rambow
- Working group chairs
- The workshop program committee
- The ACL meeting organizers
- The National Science Foundation
  (Award Nos. BCS-1404352 and IIS-1027289)