Computational Linguistics in Support of Linguistic Theory

Emily M. Bender & D. Terence Langendoen
83rd Annual Meeting of the Linguistic Society of America
San Francisco, CA
January 9, 2008

Introduction

- What can computational linguistics do for linguists?

- We aim to convince you to:

  - ... try current computational methodologies

  - ... teach your students to use computational methodologies

  - ... collaborate in building the next generation infrastructure for computational methods in linguistics

Overview

- What computers can do for us
- What infrastructure already exists
- How the next generation infrastructure will be built
- What we need to do
- Slides and links to cited projects available here:
  http://faculty.washington.edu/ebender/LSA2009.html

Overview

- What computers can do for us
- What infrastructure already exists
- How the next generation infrastructure will be built
- What we need to do

What computers can do for us

- Computational methodology can allow linguists to:
    - Access/manipulate more data
    - Collaborate with more people across greater distance
    - Ask questions not previously askable
- Here are some examples...

Descriptive & documentary linguistics

- Given a transcribed and translated narrative, which words are likely to belong to the same lemma?
- Given a partial morphological analysis, which words in the text are still unknown?

EARL (Moon & Erk 2008, Palmer & Erk 2007)

- Given data in an IPA transcription, which phones are likely allophones? What are some likely phonological rules?

(Farrar & Hou, in progress)

- What are potential cognates for these words in related languages, transcribed with different transcription systems?

OATS (Moran, in progress)

Phonetics & Phonology

- How do different feature systems quantify the variation across languages differently?

- Which feature systems locate differences in historically plausible ways, such that differences among historically or areally related languages are less pronounced?

PHOIBLE (Moran & Wright 2009)

- Given a set of OT constraints, what is the range of languages predicted?

Eraculator (Riggle et al 2007)

- What kind of data is required for learning rankings of a given set of OT constraints?

(Boersma & Hayes 2001)

Morphosyntax

- Which languages have ergative-absolutive case marking and object agreement on the verb?

- Which languages have anti-passive voice and reflexives expressed through affixes?

ODIN (Lewis 2006), WALS (Haspelmath et al 2008)

- How does my new analysis of X interact with the rest of the grammar?

- How many analyses does my grammar assign to this sentence?

- How many realizations does my grammar assign to this input semantics?

Grammar engineering (Butt et al 1999, Copestake 2002, Baldridge et al 2007, Crabbé & Duchier 2005, Bateman 1997, ...)

Semantics & Pragmatics

- What proportion of agents are non-animate, across languages and across grammatical functions?

FrameNet (Atkins et al 2003)

- What resources do languages use to express spatial, temporal, modal, evidential, ... [fill in your favorite semantic category here] properties and relations?

TimeML (Pustejovsky et al 2005)

- What words are used as exclamatives, and in what dimensions do they differ from each other?

(Potts & Schwarz 2009)

Psycholinguistics & Language Acquisition

- Do children acquiring different languages expand their function morpheme vocabulary at the same rate?
- How does the course of acquiring the restrictions on the binding of anaphors proceed across languages? What pattern of “errors” are observed?

CHILDES and TalkBank (MacWhinney 2000)

- How are speaker’s choices in utterance generation influence by various factors such as information density?

(Jaeger et al 2009)

- Are speakers sensitive to the frequency of four-word phrases, once we control for the frequency of their sub-parts?

(Arnon & Snider 2009)

Language Variation & Change

- How does lexical frequency interact with other internal constraints on sociolinguistic variation?

(Oxley, to appear)

- I have a frequency analysis of X in dialect Y. Do other varieties show similar patterns of variation?

- How much contact was there among the various branches of Indo-European early in their history?

(Nakhleh, Ringe, & Warnow 2005)

- How did *do* support spread across grammatical contexts in the history of English?

(Han & Kroch, 2000)

- Which, if any, properties of construction X cluster together crosslinguistically?

(Bickel 2007)

Summary

- Across all of the subfields of linguistics, computational methods can take us to the next level
    - Work with more data:
        - Annotate more data, more efficiently
        - Through machine-mediated collaboration, construct larger, more cross-linguistic datasets
        - With machine assistance, systematically handle more data than otherwise possible
- Test the interaction of formal rules in the complex systems we model

Overview

- What computers can do for us
- What infrastructure already exists
- How the next generation infrastructure will be built
- What we need to do

Annotation systems and standards

- Each level of analysis in linguistics relies on previous analyses, down to phonetic transcription

- Annotation systems encode analyses at one level so they can be used as data in the next

- It is tempting to use tacit speaker knowledge to skip over some of the levels

- ... but this doesn’t scale, and it can be error-prone

Annotation standards

- We need a method of representing analysis that:
    - we trust
    - is robust across languages and theories
    - will scale to many kinds of use
- The richer the structure of the data, the more interesting the questions you can ask
- But it has to be consistent: Are we building a house of cards, or something that will support the weight of further analysis?
- To scale up, so others can use our data, so we can work with more data: Annotation standards

Annotation standards: We have some!

- Segment encoding: IPA (International Phonetic Association 1999), and more generally, Unicode (The Unicode Consortium 2007)
- Prosody and intonation: ToBI (Silverman et al 1992)
- Sublexical annotation: Leipzig Glossing Rules for interlinear glossed text (IGT; Haspelmath 2003); E-MELD “best practice” recommendations (Bow, Hughes & Bird 2003)
- Supralexical annotation (Treebanks):
  - Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993)
  - Unified Linguistic Annotation (ULA) effort (for integrating documents using Penn Treebank, PropBank, NomBank, TimeML, etc. standards; Verhagen, Stubbs & Pustejovsky 2007).
  - Cross-linguistic treebank standards under development.
- Resource discovery: OLAC metadata standard (Bird & Simons 2001)

Existing data collections

- LDC: Clearing house for corpora (raw and annotated)
- CHILDES: Database of child and caregiver speech
- WordNet: Lexical resource structured around synonymy
- FrameNet: Text annotated with semantic frames and roles
- WALS: Typological properties of 2,650 languages
- Ethnologue
- ODIN: On-line Database of INterlinear glossed text
- PHOIBLE (and SOWL): Databases of phonetic segment inventories

Rise and role of documentary linguistics

- Documentary linguistics is a relatively recent development in the field, largely responsive to the awareness that the world's languages are disappearing faster than we can marshal the resources to analyze them.

- "The product [of documentary linguistics] is the primary data — a corpus of recorded speech events that document the language in actual use." It complements descriptive and analytical corpora, and can be developed in far less time than richly annotated language corpora, and if properly archived and access privileges provided, can be analyzed and annotated later.

(Simons 2008)

Overview

- What computers can do for us
- What infrastructure already exists
- How the next generation infrastructure will be built
- What we need to do

A linguistics research environment of the future

- A web service through which you could access and interact with on your own or with partners around the world:
    - analyzed, annotated texts and examples in all the world's languages
    - including child language
    - associated with sound and video files
    - as well as quantitative data from psycholinguistic experiments and typological information
    - ... which is searchable by language, linguistic feature, geographical region
- What is the minimum amount of data you'd like to see on each language in such a system?
- How useful would it be if we got even only part way there?

Such environments exist today in other fields

- Biochemistry: The Protein Folding Database

  http://pfd.med.monash.edu.au/public_html/index.php

- Nanotechnology: Nanomaterial Database

  http://www.nanowerk.com/

- Astronomy: National Virtual Observatory

  http://www.us-vo.org/

Integrated perspectives

- “All astronomers observe the same sky, but with different techniques, from the ground and from space, each showing different facets of the Universe. The result is a plurality of disciplines (e.g., radio, optical or X-ray astronomy and computational theory), all producing large volumes of digital data. The opportunities for new discoveries are greatest in the comparison and combination of data from different parts of the spectrum, from different telescopes and archives.”

“Harnessing the Power of Digital Information for Science and Society”, Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, to appear

Identified needs (1)

- Interoperability, including:
    - Standards: Data types, annotation standards, ontology
    - Tools: Software assisting ordinary working linguists in what they want to do, which as a side effect promotes the use of standards
- Data reliability:
    - Verification (where did this data come from, and how can I be sure it’s reliable?)
    - Validation (is this data set consistent internally, and with other information I have?)
- Provenance: A system for giving credit for original contributions of data

Identified needs (2)

- Application Programming Interfaces (APIs) to linguistic databases:
    - Allow developers to create applications that draw on linguistic data
    - Allow access to the data for aggregator applications which support analyses across datasets
- An organizational structure for overseeing the development, implementation, and maintenance of the resource
    - The organization needs an economically sustainable model not (solely) dependent on grants in the long run (though it could be initially started using grant funding)
    - Also needs buy-in from the research community from the outset. “Build it and they will come” probably won’t work.

How can we do it?

- High-profile workshops to get the research community's input.
- Create a "virtual organization" that encourages participants to contribute ideas, data, software, etc. without necessarily having to meet, but also to meet formally or informally at conferences and other venues.
- Cast a wide net to encourage participation. Find partners in:
  - scholarly organizations (LSA, ACL, counterparts in other countries)
  - industries with a commercial interest in language data and linguistic analysis of those data (both big, like Microsoft, Google, Nokia and IBM, and small)
  - government agencies with interest in such data and analysis
  - private individuals with an interest in language, both as data providers and data users ("citizen scientists")

Overview

- What computers can do for us
- What infrastructure already exists
- How the next generation infrastructure will be built
- What we need to do

What linguists can do

- In addition to participating in the process, there are three main things linguists can do:
    - Share data
    - Teach
    - Effect culture change

Sharing data

- Use the standards that are available
    - ... and if you don’t like them, give feedback to the relevant working groups, or join them!
- Seek appropriate human subjects permissions to make data distributable
- Publish data sets electronically
    - ... in existing repositories, or independently (see Simons 2008 for suggestions), but marked up with OLAC metadata for discoverability

Teaching: What all linguists should know

- What resources exist
- What standards/best practices exist (UNICODE, how to enter IPA...)
- Basic corpus manipulation tools (e.g., grep and wc in unix)
- Basic database querying techniques (SQL)
- How to access grid computing (SIDGRID, Teragrid)
- Sub-field specific high-level "linguistic programming" languages and/or statistics software
- Associated skills: version control, debugging, regression testing
- How to learn more about programming, if desired

Culture change

- Establish a culture of giving academic credit for creating/curating data sets
- Establish a culture (as reviewers) of expecting data sets to be published
- Establish a culture (as reviewers) of expecting claims to be tested against web-available data

Conclusion

- We wanted to convince you to:
    - ... try current computational methodologies
    - ... teach your students to use computational methodologies
    - ... collaborate in building the next generation infrastructure for computational methods in linguistics

Conclusion

- We've described a vision of cyber-enabled linguistics.
- To get there, we need to:
    - Build infrastructure, including standards
    - Contribute data
    - Promote and expect wide-spread use of the infrastructure

To learn more: Session 30, tomorrow morning

- Bender: Computer-assisted syntactic analysis; validating syntactic analyses against data
- Baldridge, Erk, Moon & Palmer, A.: Machine learning techniques from computational linguistics to reduce the cost of producing IGT in language documentation
- Xue, Brown, & Palmer, M.: Data-driven and theory-driven approaches to large-scale annotation, and their impact on theory and applications
- Riggle & Goldsmith: Information theoretic comparison of phonological models, measurement of the effect of adding structure
- Potts & Schwarz: Corpus-based approach to identifying exclamatives, measuring their emotive content, and exploring the phenomenon of exclamativity

Thank you!