Computational Linguistics for Low-Resource Languages

November 23, 2011

Alexis Palmer

Nachrichten/News

- ✦ wiki.coli.uni-saarland.de/cl4lrl/main -- new subject-area pages for notes, links, other useful information
- ✦ reviews -- consider schedule, interests, etc.; reviews to be submitted *before* class
- ✦ next week:
  - Poornima & Good 2010
  - Lewis & Xia 2010

Today

- Human Language Project (Abney & Bird 2010)
- Data model for Universal Corpus (Abney & Bird 2011)
- Data portability (Bird & Simons 2003)
- ... these slides are to guide our discussion.

The Human Language Project: Building a Universal Corpus of the World's Languages

Steven Abney and Steven Bird

ACL 2010, position/challenge paper

Motivations

Urgency

Motivations

Urgency

- critical period is NOW
- expertise in computational linguistics needed

> "The next generation will forgive us for the most egregious shortcomings in theory construction and technology development, but they will not forgive us if we fail to preserve vanishing primary language data in a form that enables future research."

Motivations

Urgency

- critical period is NOW
- expertise in computational linguistics needed

> "The next generation will forgive us for the most egregious shortcomings in theory construction and technology development, but they will not forgive us if we fail to preserve vanishing primary language data in a form that enables future research."

In the name of linguistics (and CL)

Motivations

Urgency

- critical period is NOW
- expertise in computational linguistics needed

> "The next generation will forgive us for the most egregious shortcomings in theory construction and technology development, but they will not forgive us if we fail to preserve vanishing primary language data in a form that enables future research."

In the name of linguistics (and CL)

- "complete digitization of every human language"
- universal theory/understanding of human language
- enable data-lean approaches via cross-linguistic modeling
- support automatic processing for all (?) languages

Guiding principles

Important foundational principles/goals

- Universality
- Machine readability and consistency
- Community effort
- Availability
- Utility
- Centrality of primary data

What to include -- 1 of 2

Essential components of the Universal Corpus

- Metadata
- Written text:
    - primary data
    - transcriptions
- Spoken text:
    - audio recordings
    - audio transcriptions
    - written transcriptions

What to include -- 2 of 2

Essential components of the Universal Corpus

- Both written and spoken text:
  - translations into reference language
  - sentence-level segmentation & translation
  - word-level segmentation & glossing
  - morpheme-level segmentation & glossing
- Secondary resources
  - lexicon with glosses in reference language
  - paradigms and phonological information sufficient to build a morphological analyzer

Interlinear glossed text (IGT)

Producing interlinear glossed text (IGT) starts from:

- Transcription of recorded speech
- Translation of transcribed text (may be literal or free)

(a) **xelch li.**
(b) ***Salio entonces.***
(b′) *Then he left.*

Interlinear glossed text (IGT)

Producing interlinear glossed text (IGT) involves:

- morphological segmentation
- stem translation
- morpheme glossing

(a) xelch li.
(b) x- el -ch li
(c) COM- salir -DIR DEM

(d) *Salio entonces.*

Interlinear glossed text (IGT)

Producing interlinear glossed text (IGT) involves:

- morphological segmentation
- stem translation
- morpheme glossing
- POS-tagging of stems (often derived automatically from lexicon)

```
(a)    xelch                    li.
(b)    x-      el       -ch     li
(c)    COM-    salir    -DIR    DEM
(d)    TAM-    VI       -DIR    PART
(e)    Salio entonces.
```

Interlinear glossed text (IGT)

Producing interlinear glossed text (IGT) involves:

- morphological segmentation
- stem translation
- morpheme glossing
- POS-tagging of stems (often derived automatically from lexicon)

```
(a)    xelch                    li.
(b)    x-      el       -ch     li
(c)    COM-    salir    -DIR    DEM
(d)    TAM-    VI       -DIR    PART
(e)    Salio entonces.
```

What's wrong with current resources?

Problems with some extant text collections

- Traditional language archives
  - potentially broad coverage *but*
  - restricted access
  - not amenable to machine processing
- Large-scale data collection efforts
  - e.g. LDC (Linguistic Data Consortium), ELRA (Europear Language Resources Association)
  - minimal coverage
- General problems
  - discoverability
  - standardization

The Universal Corpus ecosystem

Roles

- Editors
- CL research
- Tool builders
- Volunteer annotators
- Documentary linguists
- Data agencies
- Language archives
- Funding agencies

Where do we begin?

Early tasks

- Seed corpus
- Resource discovery
- Resource classification
- Acquisition
- Text collection
- Audio protocol
- Corpus readers

Outgrowths

Already in progress

- ✦ BOLD:PNG and larger (100-lg) project
- ✦ All Languages Wiki (alpha version)
- ✦ Language Commons portal
- ✦ OLAC

Seven Dimensions of Portability for Language Documentation and Description

Steven Bird and Gary Simons
Language 2003

The context

Foundational paper for digital documentation/description

- Language documentation v. language description
- Dated in some ways, utterly relevant in others
- Tools and technologies
    - General purpose tools
    - Specialized tools
    - Digital technologies
    - Digital archives

The seven dimensions

Content

- Coverage
- Accountability
- Terminology

Format

- Openness
- Encoding
- Markup
- Rendering

The seven dimensions

Discovery

- Existence
- Relevance

Access

- Scope of access
- Process for access
- Ease of access

The seven dimensions

Citation

- Bibliography
- Persistence
- Immutability
- Granularity

Preservation

- Longevity
- Safety
- Media

The seven dimensions

Rights

- Terms of use
- Benefit
- Sensitivity
- Balance

Impact

Many of the recommendations have become standards

- XML
- Unicode
- Archiving and metadata standards (OLAC)
- Citation persistence
- Open source