Computational Linguistics for Low-Resource Languages

4 May 2016
Alexis Palmer
palmer@cl.uni-heidelberg.de
• topics and schedule posted on Thursday or Friday
• May 11: LRAs from Mayumi Ohta, Juri Opitz
• May 11: active learning for morpheme glossing

• don’t forget to send questions
The Human Language Project: Building a Universal Corpus of the World’s Languages

Steven Abney and Steven Bird
ACL 2010, position/challenge paper
Aim

- corpus of world’s languages: cf. HGP
- consistent, machine-readable format
- fixed set of data types
- enable massive cross-linguistic processing

<table>
<thead>
<tr>
<th></th>
<th>Ainu</th>
<th>Nahuatl</th>
<th>German</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bitext</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swadesh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Human Language Project - the vision

Aim

- corpus of world’s languages: cf. HGP
- consistent, machine-readable format
- fixed set of data types
- enable massive cross-linguistic processing
Motivations

Urgency
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
What counts as sufficient digitization of a language?

• one view: enough to enable language learning in the absence of native speakers

• another view: enough to enable linguistic analysis in the absence of native speakers

• yet another view: enough to provide an appropriate cultural record (moving target...)

Digitizing a language
Abney and Bird’s view:

- use case: machine translation, data-lean methods
- MT as proxy for language learning, semantic interpretation
- for each language: primary texts w/ translation to reference language
Digitizing a language

Anticipated “pyramidal” structure of resources

unannotated text
translation from target lg. to reference lg.
sentence and word alignments
morphological analysis
interlinear glossed text
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.
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.
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.
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.
Terminology and background

- Metadata - Dublin Core, TEI
- Provenance
- Source citation (beyond bib. entry)
- Spontaneous vs. elicited speech
- Slow speech “audio transcriptions”
- Orthographic vs. phonetic transcription
- Attested forms
- ISO 639 language codes
Dublin Core Metadata Elements

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights
Essential components of the Universal Corpus

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

ID: europarl/swedish/ep-00-01-17/18
LANGS: swd eng
SENT: det gäller en ordningsfråga
TRANS: this is a point of order
ALIGN: 1-1 2-2 3-3 4-4 4-5 4-6
PROVENANCE: pharaoh-v1.2, ...
REV: 8947 2010-05-02 10:35:06 leobfld12
RIGHTS: Copyright (C) 2010 Uni...; CC-BY
<table>
<thead>
<tr>
<th>ID: example/001</th>
<th>ID: swadesh/47</th>
</tr>
</thead>
<tbody>
<tr>
<td>LANGS: eng</td>
<td>LANGS: fra</td>
</tr>
<tr>
<td>SENT: the dogs are barking</td>
<td>LEX: chien</td>
</tr>
<tr>
<td>LEX: the dog be bark</td>
<td></td>
</tr>
<tr>
<td>AFF: - PL PL ING</td>
<td>ID: swadesh/47</td>
</tr>
<tr>
<td></td>
<td>LANGS: eng</td>
</tr>
<tr>
<td></td>
<td>LEX: dog</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID: swedishlex/v3.2/0419</th>
<th>LANGS: swd eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEX: ordningsfråga</td>
<td>TRANS: point of order</td>
</tr>
</tbody>
</table>
Guiding principles

**Important foundational principles/goals**

- Universality
- Machine readability and consistency
- Community effort
- Availability
- Utility - lightweight vs. rich annotation
- Centrality of primary data - analyses ideally reference primary data
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
Potential research directions

- Active learning to improve quality of alignments and bilingual lexicons
- Automatic language identification for low-resource languages
- Automatic morphology learning for low-resource languages
- Leaner methods for building MT systems
- New algorithms for cross-linguistic bootstrapping via multiple paths
- More effective techniques for leveraging human effort
- …
Potential development directions

- Tools for various stages in the process
- Pilot studies to determine documentation cost for different language categories
- Development of corpus readers, format converters, etc.
Outgrowths

Already in progress

- **BOLD:PNG** and larger (100-lg) project
- **All Languages Wiki** (alpha version)
- **Language Commons** portal
- **OLAC**
- **Alliance for Linguistic Diversity:** endangeredlanguages.com

- Relevant open-source code: [https://github.com/RichardLitt/endangered-languages](https://github.com/RichardLitt/endangered-languages)
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

• Sidetrip: Summer Institute for Linguistics (SIL)
Tools and technologies

• General purpose tools
  • office software - WYSIWYG
  • hypertext
  • database packages
  • presentation vs. underlying data structure

• Specialized tools
  • Shoebox/Toolbox, Fieldworks Language Explorer
  • Praat, ELAN, etc.

• Digital technologies
• Digital archives
Tools and technologies

- General purpose tools
  - office software - WYSIWYG
  - hypertext
  - database packages
  - presentation vs. underlying data structure
- Specialized tools
  - Shoebox/Toolbox, Fieldworks Language Explorer
  - Praat, ELAN, etc.
- Digital technologies
- Digital archives

“digital detritus”
The seven dimensions

Content

• Coverage - weaknesses in coverage limit interpretability of data
• Accountability - allow reference to original linguistic event
• Terminology - GOLD

General Ontology for Linguistic Description (Farrar & Lewis 2007)

```
Thing
  \_ Abstract
    \_ Linguistic Property
      \_ Morphosemantic Property
        \_ Tense Property
        \_ Future Tense
```
Examples

NO_COMMENT
2009-06-04 13:28:07

Gilaabur- u- nfermahamišaluggünaamuq'- da- c
now.they- OBL- GEN.farm.forever.behind.stay- FUT- NEG

Now their farm will not stay behind forever.

References:
Haspelmath (1993:207)

This example demonstrates epistemic necessity, also known as necessary truth or presupposed truth, in that this statement presupposes that the pig will run out.

2009-06-04 13:28:07

ho bu- busal- eb age qo- qag- an
pig SIM- run.out- 3.SG.DIFSBJ.IRREAL 3.PL hit- 3.PL.FUT

They will kill the pig as it runs out.

References:
Palmer (2001:5) FROM Roberts (1990)

linguistics-ontology.org
The seven dimensions

Format

• Openness - the problem of proprietary formats
• Encoding - despite significant work in this area, a perpetual problem when working with language data
• Markup - need for machine-readable markup
• Rendering - need for human-accessible display

(1) a. chien n dog.
    b. chien: [n] dog.

(2) \ent chien
    \pos n
    \def dog
The seven dimensions

a. <p><font size=+1><i>chien</i></font><b>n</b><font color=blue>dog.</font></p>

b. <entry>
<headword>chien</headword>
<pos>n</pos>
<definition>dog</definition>
</entry>
The seven dimensions

Discovery
- Existence
- Relevance - role of controlled vocabulary

Access
- Scope of access
- Process for access
- Ease of access - importance of printed versions, access via DVD/etc. (i.e. without internet connection), importance of both online and offline tool versions
The seven dimensions

Citation
- Bibliography
- Persistence
- Immutability/versioning
- Granularity - locating the precise reference-instance/lexical item

Preservation
- Longevity
- Safety
- Media
The seven dimensions

Rights

- Terms of use - complex ethical issues
- Benefit
- Sensitivity - different perspectives on data
- Balance
Many of the recommendations have become standards

- XML
- Unicode
- Archiving and metadata standards (OLAC)
- Citation persistence
- Open source
- Movement within ACL/LSA/etc. to encourage (sometimes even require) submission of data with publications
Questions from you

• rights re: crawled data?
• HLP: list of numbers given as one way of representing data - what do these mean? [9347, 3053, 0038, …]
• what about all of the pragmatic and world knowledge that gets lost if we consider translation-capability a measure of an adequate digitization?
• is primary text always written in Latin letters? what about other/no writing systems?