Course requirements & organization

- **course website:**
  www.cl.uni-heidelberg.de/courses/ss16/cllrl/

- **schedule and literature to be posted on course website**

- **your slides will also be posted**

- **language:** auf Deutsch geht auch
Palmer, ICL, UHeidelberg

Course requirements & organization

✧ **reading & participation**: read papers prior to relevant meeting, discuss

✧ **questions**: 2 questions/session, submitted (email) *before noon* on day of class

✧ **presentation**: presentation of selected paper(s), discussion after

✧ **language resource assessment**

✧ **term paper**: original research or in-depth survey and analysis (12-15 pages)
Student presentations

- **topic**: 1-2 related papers, depending on length and complexity
- **presentation**: scheduling TBD (depends on number of students), roughly 45 minutes for presentation plus discussion
- **preparation**: draft of slides at least one week prior to presentation, meeting for feedback
- **Sprechstunde**: Wednesdays 11:30-12:30, or by appointment (M/W/Th)
Language resource assessment

- **goal**: determine the state of language resources for a language of your choice
- **presentation**: short presentation (~10 min.), schedule TBD
- **investigate**: digital language resources, any NLP tools? corpora? work on revitalization/preservation? availability of resources?
- **TODO**: choose your language before 04.05 (email me - first come, first served)
Questions of interest

- What is a low-resource language? (aka less-studied language, resource-poor language, minority language, less-privileged language, ...)

CL for LRL

Questions of interest

• What is a low-resource language? (aka less-studied language, resource-poor language, minority language, less-privileged language, ...)

• What are the challenges posed by LRLs, and what are the major approaches to addressing these challenges?
Questions of interest

• What is a low-resource language? (aka less-studied language, resource-poor language, minority language, less-privileged language, ...)
• What are the challenges posed by LRLs, and what are the major approaches to addressing these challenges?

Some major themes

• Role of labeled/annotated data
• Role of expert/linguistic knowledge (anno & beyond)
• Single language vs. “universal” solutions
• Resource creation: does it always make sense? how can it be done most efficiently?
And another question...

Why do we care?
Why do we care?

✦ practical reasons
✦ cultural reasons
✦ theoretical reasons
Language endangerment

Language loss

- Current estimated rate of language death: one every 2 weeks (Crystal 2000)
- Half of world’s languages extinct by end this century
- UNESCO Endangered Languages Programme (under auspices of Section on Intangible Cultural Heritage)
- UN General Assembly: 2008 was International Year of Languages
Language endangerment

Language loss

- Current estimated rate of language death: one every 2 weeks (Crystal 2000)
- Half of world’s languages extinct by end this century
- UNESCO Endangered Languages Programme (under auspices of Section on Intangible Cultural Heritage)
- UN General Assembly: 2008 was International Year of Languages

UNESCO endangerment status

- six levels: safe, unsafe (or vulnerable), definitively endangered, severely endangered, critically endangered
- criteria go beyond number of speakers
Evaluating language endangerment

Criteria to consider (UNESCO 2003)

• Intergenerational language transmission
• Absolute number of speakers
• Proportion of speakers within the total population
• Trends in existing language domains
• Response to new domains and media
• Materials for language education and literacy
• Governmental and institutional attitudes and policies, including official status and use
• Community members’ attitudes toward their own language
• Amount and quality of documentation
Globally, 2488 languages in danger

528 ‘severely endangered’ languages

Germany: 13 endangered languages

Documenting endangered languages

The realities

• Most projects are individual or small-group endeavors with very small budgets
• Each project seems to find its own workflow
• Basic approach: collection, transcription, translation, detailed linguistic annotation (NOT a pipeline)
• Tangible end products: orthographies, grammars, dictionaries, language teaching and learning materials, collections of stories, websites, etc.
• Such materials support survival of the language
• Do they support CL/NLP???
Uspanteko: 1320 speakers, ‘unsafe’ status

- Uspantán, Quiché Department, Guatemala
Corpus of texts in the Mayan language Uspanteko

- Produced by OKMA (Oxlajuuj Keej Maya' Ajtz'ib')
- 66 texts, mostly oral history, personal experience, and stories
- Total 284K words of transcribed text, 74K words glossed

IGT-XML: representational format specifically for IGT

<table>
<thead>
<tr>
<th></th>
<th># texts</th>
<th># morphemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>21</td>
<td>38802</td>
</tr>
<tr>
<td>dev</td>
<td>5</td>
<td>16792</td>
</tr>
<tr>
<td>test</td>
<td>6</td>
<td>18704</td>
</tr>
</tbody>
</table>
Types of resources

Data

• primary: audio, video, texts (archiving)
• machine-readable corpora
• data with annotations
• parallel corpora, comparable corpora
Types of resources

**Data**
- primary: audio, video, texts (archiving)
- machine-readable corpora
- data with annotations
- parallel corpora, comparable corpora

**Linguistic resources**
- traditional: grammars, dictionaries, word lists
- WordNet, other ontological resources
- treebanks, etc.
Types of resources

Data
• primary: audio, video, texts (archiving)
• machine-readable corpora
• data with annotations
• parallel corpora, comparable corpora

Linguistic resources
• traditional: grammars, dictionaries, word lists
• WordNet, other ontological resources
• treebanks, etc.

Tools
• user-oriented: spell checkers, input systems, etc.
• for NLP: tokenization, POS tagging, parsing, etc.
Challenges and approaches

Having to do with insufficiency of data
• create more data?
• leverage resource-rich languages
• use semi- or unsupervised methods
• use rule-based methods
• ...

Having to do with the nature of the data
• use linguistic knowledge to seed unsupervised models
• use linguistic knowledge to adapt models/approaches
• change the data to look more like familiar languages
• ...
Topics and scheduling
Topics

• More complete list of topics & readings on website
• Some options
  ‣ Data/resource creation
  ‣ POS tagging and morphological analysis
  ‣ Syntactic analysis
  ‣ Linguistic universals, linguistic typology
  ‣ Speech tools for LRLs
  ‣ Machine translation
  ‣ Cross-lingual approaches
  ‣ ...

Scheduling

- 4 May: foundations, Bird/Simons, Bird/Abney [me]
- 11 May: possible start of student presentations

For next week:
- Bird and Abney on building a Universal Corpus
- Bird and Simons on requirements for good data
- Email me with topic preferences (top 3) - by Monday (02.05)