The Linguistic Data Consortium: Developing and Distributing Language Resources for All

Denise DiPersio, Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania
{dipersio, ccieri} AT ldc.upenn.edu
Overview

- LDC: Founding and Mission
- Sharing, Curating Language Data
- Language Resource Overview
- Research Collaborations in Indigenous Languages
- Conclusion
A mutual aid society with the mission to develop and distribute language resources to the global community
- Academia, government, industry
- Researchers contribute data sets: visibility, community recognition, uptake
- Members/data licensees contribute fees: ongoing rights to a variety of resources
- Sponsors contribute funding: resource creation, infrastructure, innovation, cost sharing, resource dissemination to the community

LDC’s online Catalog launched in 1993
- Close to 200,000 copies of 820+ resources in more than 90 languages distributed to roughly 6000 distinct organizations in over 100 countries
- 3-4 new data sets released monthly
- Distributed under a variety of licensing arrangements: for use in language-related research, education and technology development

Research impact: more than 10,000 papers cite LDC data
The LDC Catalog is a permanent language resource archive

- Seeded by data contributions of significant corpora, augmented by data sets developed by LDC in funded projects along with contributions from the global research community

The Catalog is a CoreTrustSeal trustworthy repository

- Meets high standards for data access, metadata, rights management, curation, storage, security

Curation workflow: data review, quality checks, metadata, documentation

- Storage and back-up system; migration to new formats, storage, media as needed
- Licenses consistent with community use and address human subjects, privacy, intellectual property, tribal rights to community languages

LDC has the expertise and infrastructure to ensure that data is preserved and accessible, with appropriate protections to language communities, students, scholars, researchers and developers
More resources in a growing number of languages: indigenous languages, minority languages, endangered languages, low resource languages

- All are underserved language communities
- Human language technologies need digital resources
- Scarce source data, language structure present research challenges

LDC data set and research case studies

- West African languages
  - Manding and Yoruba lexicons, Dschang and Ngomba (Bantu) tone paradigms
- Fieldwork
  - Language preservation in Papua New Guinea, Brazil
  - Malto Speech and Transcripts
- Language Packs
  - Core resources and tools
<table>
<thead>
<tr>
<th>w</th>
<th>Headword</th>
<th>á</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>tone</td>
<td>H</td>
</tr>
<tr>
<td>c</td>
<td>category</td>
<td>Interj</td>
</tr>
<tr>
<td>d</td>
<td>definition</td>
<td>ah! exclamation</td>
</tr>
<tr>
<td>dfr</td>
<td>definition fr</td>
<td>ha! exclamation</td>
</tr>
<tr>
<td>e</td>
<td>example</td>
<td>Á, à má nà tůgůn.</td>
</tr>
<tr>
<td>g</td>
<td>glose</td>
<td>Ah, he/she never came back.</td>
</tr>
<tr>
<td>gfr</td>
<td>glose fr</td>
<td>Ha, il/elle n’est plus revenu/revenue.</td>
</tr>
<tr>
<td>c</td>
<td>category</td>
<td>Pro</td>
</tr>
<tr>
<td>d</td>
<td>definition</td>
<td>you, short form used in the imperative [2d pers. pl.]</td>
</tr>
<tr>
<td>dfr</td>
<td>definition fr</td>
<td>vous, forme tronquée utilisée à l’impératif [2ème pers. pl.]</td>
</tr>
<tr>
<td>e</td>
<td>example</td>
<td>Á yé tágá!</td>
</tr>
<tr>
<td>g</td>
<td>glose</td>
<td>You go!</td>
</tr>
<tr>
<td>gfr</td>
<td>glose fr</td>
<td>Allez-y!</td>
</tr>
<tr>
<td>id</td>
<td></td>
<td>300</td>
</tr>
</tbody>
</table>
Collaborative Transcription in Papua New Guinea
Language Packs

- REFLEX, LORELEI US projects
- Resources and tools
  - Monolingual, parallel text
  - Annotation
  - Tools for text processing, segmentation, entity tagging
  - Lexicons, grammatical sketches
- Multiple purposes:
  - Language documentation, preservation
  - Basic technology development
  - Situational awareness
- Akan (Twi), Amazigh, Amharic, Ilocano, Kinyarwanda, Odia, Oromo, Sinhala, Tigrinya, Uighur, Wolof, Zulu +
- In LDC catalog -- 2020
Research Collaborations in Indigenous Languages

- **Language documentation support**
  - AARDVARC (Automatically Annotated Repository of Digital Audio and Video Resources Community)
  - EMELD (Electronic Metastructure for Endangered Languages Data)
- **Advice and technical assistance for collections:** Nahuatl, Mixtec, Tembé and Nhengatu
- **LDC workshops around languages in the Americas**
  - Philadelphia 2018: Planning Workshop on Data Archives and Languages of the Americas
    - Experts managing linguistic data archives and resource centers discussing challenges, needs and opportunities for promoting and extending collaboration in the Americas
  - Mexico City 2018: International Workshop on Data Intensive Research on Languages of the Americas
    - Linguists and scientists from Mexico, Brazil, Chile, Argentina, USA
    - Languages discussed include Chuj, Yucateco, Huasteco, Nahuatl, Wixarika, Southern Cone languages, Mexican/American Spanish, Brazilian Portuguese
LDC Global Network of select data sources including: ■ = subcontractors and vendors, ● = corpus authors, ◆ = media providers, ◇ = LDC staff collections, ★ = research collaborators. Many markers represent multiple collaborators; many markers partially obscured by others.
Access: crucial theme of this International Year of Indigenous Languages
- Education, information, knowledge

Sharing data, developing language technologies echo the theme
- LDC’s founding principle: broad access to data drives knowledge and research

LDC is committed to developing and sharing resources in all languages for all language communities in ways that ensure meaningful access, advance language vitality and promote preservation