Linguistic Data Consortium

Introduction to LDC: 25 Years and Counting

Denise DiPersio, Associate Director

dipersio@ldc.upenn.edu

Overview

- The Consortium Model
- Chinese Resources
- Benefits of Sharing Data through LDC
- Innovations in Data Access and Delivery
- Current Projects/Collaborations
- 25th Anniversary Timeline

About LDC

- The Linguistic Data Consortium (LDC) is a non-profit, hosted at the University of Pennsylvania, that creates and distributes language data to universities, laboratories, companies and libraries for language-related education, research and technology development
- LDC's Catalog has over 700 holdings in more than 90 languages. Three to four data sets – text, speech, video and lexicons – are released monthly
- Current activities include:
  - Data collection and annotation
  - Software development
  - Research and project support
  - Archiving, publishing and distributing corpora

The Consortium Model

- LDC was founded to address the data shortage facing language technology research and development
  - A mutual aid society: the community supports Consortium activities that in turn benefit the community through access to data for language-related research, education and technology development

- The community supports LDC activities through:
  - Membership
    - Ongoing rights to corpora released in membership year; other data sets accessible at reduced fees
    - Access across the organization
  - Data Contributions
    - External contributions make up over 50% of the LDC Catalog's holdings
    - Source data providers also contribute (newswire, broadcast, etc.)
  - Collaborations: research, infrastructure, community networks

LDC's Global Network

corpora=red, media=purple, employees=blue, research collaborators=orange, collection/annotation subs=green

Chinese Resources Snapshot 1/2

- 170 Chinese releases (October 2017)
  - Broadcast audio and video: news and conversation
    - Broadcast speech and transcripts, TRECVid Keyframes and Transcripts, Mandarin Chinese Phonetic Segmentation and Tone
  - Conversational telephone speech
    - CALLHOME, CALLFRIEND, HUB5 evaluation data, Babel language pack
    - Mandarin, Min Nan, Wu, Yue
  - Microphone speech
    - Mandarin-English Code-Switching in Southeast Asia
  - Newswire, laws, journals, web text
    - Chinese Gigaword (5th Edition), NIST OpenMT Machine Translation evaluations, GALE, BOLT parallel and word-aligned data sets, Hong Kong Laws/Hansards and United Nations parallel text, Chinese Web 5-gram
  - Syntactic and semantic annotation
    - Chinese treebanks, propbank, discourse treebank; ACE (entities, information extraction, time normalization); TDT (topic detection); OntoNotes (co-reference)
  - Handwriting analysis – MADCAT Chinese Pilot Training Set

Chinese Resources Snapshot 2/2

- Donations
  - Academia Sinica/Hong Kong Polytechnic University: Tagged Chinese Gigaword, new releases to come
  - Chilin (HK): English-Chinese Parallel Sentences from Patents
  - Harbin Institute of Technology: Chinese Dependency Treebank
  - Hong Kong University of Science & Technology: parallel text, telephone speech, transcripts
  - Nanjing Normal University: Ancient Chinese Corpus
  - Xi'an Jiaotong University: Domain-Specific Hyponym Relations
  - Zhejiang University: Mandarin Affective Speech
- More donations welcome!

Benefits of Sharing Data through LDC

- Strong global network
  - Monthly newsletter announcing new publications reaches over 8000 recipients
  - Catalog is accessed every day by users from around the world
- Archiving and curation best practices followed
  - Quality checks, metadata schema, rights management, content delivery, permanent archive, secure storage
- Expertise in publishing and licensing data
  - Over 120,000 copies distributed under various license arrangements
- Non-exclusive rights to LDC
  - Consistent with mission to provide broad access to data

Innovations in Data Access and Delivery

- Reissuing legacy corpora
  - Updating encoding, formats, metadata, documentation
  - CALLHOME Mandarin speech, HUB5 Mandarin speech/transcripts
- Catalog and business system enhancements
  - Incorporates e-commerce principles
  - Users have increased control over their LDC accounts
  - Can license data and join LDC online
  - Search Catalog by macro-language and dialects, https://catalog.ldc.upenn.edu/search
- Data Delivery
  - Most LDC resources can now be downloaded electronically
    - Data sets up to 32 GB: digital delivery (from LDC, cloud, grid)
    - 32 GB – 64 GB: USB flash drive
    - 64 GB+ : Hard drive
  - Faster access; reduced reliance on shipping

LDC Projects, 1/3

Current projects in which LDC is involved include…

- LORELEI (Low Resource Languages for Emergent Incidents) (DARPA)
  - LORELEI seeks to identify the elements that different languages have in common and use that knowledge to enable rapid, low-cost development of automated language capabilities for use with low-resource languages for effective situational awareness
  - LDC supports LORELEI by collecting, creating and annotating linguistic resources in multiple languages
- DEFT (Deep Exploration and Filtering of Text) (DARPA)
  - DEFT develops automated systems to process text information and enable the understanding of connections in text not readily apparent to humans
  - LDC supports DEFT by collecting, creating and annotating a variety of Chinese, English and Spanish data sources to support Smart Filtering, Relational Analysis, Anomaly Analysis, Committed Belief

LDC Projects, 2/3

- TAC (Text Analysis Conference) KBP (NIST)
  - TAC is a series of evaluation workshops organized by NIST to encourage research in Natural Language Processing and related applications
  - LDC provides Chinese, English and Spanish linguistic resources for the KBP (Knowledge Base Population) Track, which promotes research in automated systems that can discover information about named entities as found in a large corpus and incorporate this information into a knowledge base
- NIEUW (Novel Incentives and Workflows in Linguistic Data Collection and Annotation) (NSF)
  - Building a portal infrastructure to access games with a purpose, citizen science and language professional tasks designed to develop multilingual language resources

LDC Projects, 3/3

- Language Application Grid (NSF)
  - NSF-sponsored collaboration among Vassar University, Brandeis University, Carnegie Mellon University and LDC
  - The goal is to develop a platform for natural language processing tools and resources that can be used and accessed by any researcher or developer
- Syntactic Parsing Project (completed)
  - Collaboration between LDC and Google to create parsing resources to improve syntactic web searches
  - LDC published two data sets from this effort:
    - English Web Treebank (August 2012)
    - English News Text Treebank: Penn Treebank Revised (July 2015)
- LDC also supports sponsored projects by sharing its expertise and by distributing language resources

25th Anniversary Timeline

Linguistic Data Consortium

Questions? Thank you!