Adapting to Trends in Language Resource Development:
A Progress Report on LDC Activities

Christopher Cieri, Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
{ccieri,myl} AT ldc.upenn.edu
LREC goal: understand the HLT landscape (Calzolari, opening ceremony)

Constant: demand: languages, annotation sophistication, communities

Changing: relative priority: volume, complexity, richness, multilinguality, multimodality

Some HLT approach human performance shifting focus to quality, richness over quantity

Elsewhere, data demand (supply) exceeds what was conceivable a few years ago: Gigawords corpora, Google n-gram corpora

New research communities begin to adopt corpus based methods
  - advanced practitioners blurs traditional boundaries (Yaeger-Dror 2002, Clopper & Pisoni 2006)
  - others await adaptive access to existing data and flexible standards

Worldwide spread of computing increases languages on web and consequently demand for LRs

Computing permits even solitary researchers to produce large, rich corpora

Yet demand for Data Centers continues to grow and mutate
Role of Data Centers

- Specialized Publishers
- Archives
- Creators & Validators of Databases
- Specification Writers
- Developers of Tools and Standards
- Technology Evaluators
- Project Managers
- Consultants, Trainers
Linguistic Data Consortium established 1992
- via open, competitive DARPA solicitation, won by U. Penn.
- centralize data distribution/archiving of language data, manage licenses, distribution practice
- structured as consortium, organization of organizations

Business Model
- developed by overseers from government, industry and academia
- DARPA funding covered operations, corpus creation for 5 years
- required to be self-sufficient via annual membership fees, data licenses
- grants fund LR creation, not maintenance; NSF, NIST early supporters

Data Sources
- donations, funded projects, community initiatives and LDC initiatives

Membership
- members provide annual support generally fees, sometimes data, services
- receive ongoing rights to data published in years when they support LDC
- reduced fees on older corpora, extra copies
Benefits

- Uniform licensing within & across research communities
  - 4 basic user license types, 1000s of instances
  - ~100 provider arrangements
  - no significant copyright issues in 17 years of operations
  - several independent issues resolved

- Cost Sharing
  - relieves funding agencies of distribution costs
  - provides vast amounts of data to members
    - LDC annual membership benefit ~30 corpora
    - development cost for 1 corpus ≥ (LDC membership fee * 10 | 100 | 1000)

- Stable research infrastructure
  - LRs permanently accessible
  - terms of use & distribution methods standardized & simple
  - members’ access to data ongoing
  - any patches available via same methods
  - tools, specifications, papers distributed without fee
LDC Roles

- distribution & archiving
- language resource production, including quality control
- intellectual property rights and license management
- human subject protocol management
- data collection
- annotation and lexicon building
- creation of tools, specifications, best practices
- knowledge transfer: documentation, metadata, consulting, training
- corpus creation research (meta-research) and academic publication
- resource coordination in large multisite programs
- serving multiple research communities
  - as funding panelists, workshop participants and oversight committee members.
Since inception in 1992, LDC has distributed

- >68,000 copies of
- 1,000 titles to
- 2,800 organizations in
- >65 countries

About half of the titles are e-corpora

- developed for technology evaluation program
- released generally after use in the relevant communities

63 titles added to Catalog since last LREC
Observation: three modes of use of LDC Data
- 1-3 | 12-16 | all

Adaptation: new membership models
- standard: 1 copy of ≤16 corpora, upon request, perpetual rights, reduced fees for older corpora, extra copies
- subscription: two copies, on media, all corpora released, shipped automatically
  - report greatest satisfaction rating among LDC members

Observation: miscellaneous requests for reduced fees, mostly from students

Adaptation: LDC Scholarships in Data
- LDC Principle: no one with a bona fide research agenda and a genuine lack of ability to contribute will go without data
- Scholarships @ semester, fund endowment at least equivalent to current expenditure
- Requirements: strict adherence to application requirements, data use statement, letter of support from advisor
- Primary Review by LDC staff, secondary review where needed by experts
Adaptation: Publications

- Observation: need for data in increasing variety of languages
- Adaptation:
  - ongoing relationships with providers around the world
    - W. Bohemia, West Point, Google, IIT Bombay, Lancaster, Colorado
  - Expansion of LDC’s own data production and distribution
    - Gigawords: English Chinese, Arabic, French, Spanish
    - Dictionaries: Tamil, Yoruba, Mawu
- Observation: shift in HLT activity
- Adaptation: shift in publications
  - NLP, 19 corpora
  - machine translation: 14 corpora
  - speech to text, 83% non-English
  - information extraction
  - language modeling, 8 corpora
  - language and speaker recognition
Adaptation: Programs

NSF
- SCOTUS (Supreme Court of the U.S.) – digital speech & aligned transcripts
- Digging into Data – Mining a Year of Speech

DARPA
- GALE – word level alignment, higher accuracy accurate Treebanks
  - ongoing relationships with HKUST, Med-LTC, MediaNet
- MADCAT – handwriting analysis, also in Arabic
- MR – tagging extents of mentions of ontology instances in text

NIST
- SRE – multichannel, multi-genre including
- LRE – BNBS

IARPA
- Aladdin – recognizing events in audio-visual data

DOE IRSG – updated digital dictionaries based on GUP Iraqi, Syrian, Moroccan

Phanotics – tagging socio-linguistic/dialect features for speaker/dialect recognition
Adaptation: Data Collection

- news text
- web text: newsgroups, blogs, zines
- biomedical text & abstracts
- printed, handwritten & hybrid documents
- broadcast news
- broadcast conversation
- conversational telephone speech
- lectures
- meetings
- interviews
- read & prompted speech
- role play
- web video
- animal vocalizations
Adaptation: Annotation

- data scouting, selection, triage
- audio-audio alignment; bandwidth, signal quality, language, dialect, program, speaker
- quick and careful transcription, aligned at the turn, sentence, word level
- orthographic & phonetic script normalization
- phonetic, dialect, sociolinguistic feature & supralexical
- documenting zoning
- tokenization and tagging of morphology, part-of-speech, gloss
- syntactic, semantic, discourse function, disfluency, sense disambiguation
- relevance
- identification, classification of mentions in text of entities, relations, events & coreference
- knowledgebase population
- time & location
- summarization of various lengths from 200 words down to titles
- translation, multiple translation, edit distance, translation post-editing, translation quality control
- alignment of translated text at document, sentence & word levels
- physics of gesture
- identification, classification of entities and events in video
Adaptation: Program Services

- ongoing assessments of sponsors’, developers’, evaluators’ needs
- timelines for LR creation and system evaluation
- translates underspecified “wish lists” into a feasible action plan
- coordinates LR creation activities across entire program and with other programs and funding agencies
- maintains data matrix of programs’ LR features and availability
- ongoing discussion, optimization, stabilization of data requirements
- incorporate technology into data production improving
- rapid program data cataloging licensing, replication, distribution
- broadening program impact through general distribution
- protection of restricted data
|                              | DARPA | NSF  | Early |
|------------------------------|-------|------|-------|
| Development Cost            | Sponsor | Sponsor | Sponsor |
| Internal Distribution       | Sponsor | Sponsor | User   |
| General Distribution        | User   | Sponsor | User   |
Data Centers must adapt in order to continue a central role in LR development and sharing efforts.

Data Centers must continue this role because they alone offer the:

- dedicated labor force
- specialized equipment
- special training

needed to fulfill their mission of lower barriers to LR access:

- simplify search
- guarantee longevity
- reduce cost