PROJECTS

Overview: Sponsored Projects Support

Language Resource Projects

The Linguistic Data Consortium (LDC) provides a wide range of Language Resources (LR) and services for Human Language Technology (HLT) research and development programs. These include corpora and lexicons, hosting, rights management, customized tools and specifications, needs analysis and outsourcing, products and services that have evolved as LDC supported the largest technology programs.

Complexity & Simplicity

With almost two decades of institutional experience in LR creation and distribution, LDC has developed approaches to successfully support HLT R&D programs. Today's programs are complex and rest atop sets of partially contradictory needs. The research agendas and commercial goals of technology developers align imperfectly with the needs of program managers and their constituents, the requirements of robust evaluation and the abilities of human subjects, annotators or analysts. There is natural tension among quantity, quality, time and cost. Furthermore, the needs of such programs evolve over time, generally requiring greater volume, accuracy, sophistication and compliance with standards as the technology matures.

Planning

To simplify project planning and management, LDC:

- Consults with program managers on LR needs
- Conducts initial and ongoing assessments of sponsors', developers' and evaluators' needs and negotiates timelines for LR creation and system evaluation
- Facilitates ongoing discussion, optimization and stabilization of data requirements
- Translates underspecified, often contradictory "wish lists" into a feasible action plan
- Develops and maintains a Data Matrix cataloging of all programs' LR features and availability
- Coordinates LR creation and sharing across entire program and with other programs and funding agencies
- Incorporates emergent technology into data production to improve efficiency and quality

Best Practices

The timeline for HLT programs is too short to reinvent the wheel. LDC adopts existing best practices or, where absent, establishes new ones for LR creation and dissemination. LDC documents best practices through formal task specifications, annotation guidelines and format standards.

Data Scouting and Collection

- Newswire, newspapers, web text (blogs, zines, newsgroups), biomedical texts
- Broadcast and webcast audio and video: news and conversation
- Telephone conversation, meetings, interviews, lectures
- Read and prompted speech
- Printed, handwritten and hybrid documents

Source Management

Large volume collection and annotation requires coordinating the activities of many external organizations and individuals and maintaining relationships with others. LDC works with hundreds of news agencies, broadcasters and web hosts to allow their material to be used in HLT research. In this capacity, LDC acts as an **intellectual property rights** intermediary, negotiating on behalf of all supported research communities.

LDC creates protocols and acquires **informed consent** for all data collected from **human subjects**. These protocols are of maximal generality for research projects with similar risk/benefit ratios to allow for rapid ramp-up of new collections.

When collection or annotation can be done more efficiently elsewhere, LDC **outsources** or **crowd-sources** the work to reduce cost and increase scope. Such approaches require different management strategies and experience in order to succeed. LDC has collaborated with dozens of organizations around the world including transcription and translation agencies, and commercial and academic research groups who collect and annotate on behalf of the Consortium.

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Linguistic Annotation

It has been said that "the best data is more data" but in fact the best data for HLT programs are large corpora annotated to match specific needs. Over the past decade, LDC has developed competence in a multitude of annotations applied to many languages including more than a dozen less commonly taught languages.

Infrastructure

Appropriate infrastructure increases the volume, efficiency, consistency, suitability and sophistication of LRs. LDC develops and shares collection and annotation infrastructure including collection platforms, data harvesting utilities, automated annotation processes and multilingual, multiplatform, freely distributed annotation toolkits.

Distribution

The true measure of the value of an LR is how much it is used. LDC's prime directive is to share LRs including:

- Rapid, regular cataloging, licensing, replication and distribution of program data
- Broadening program impact by distributing widely, using a variety of cost-sharing models
- Appropriate restriction of FOUO, GFI data, evaluation and progress sets

Research

At LDC, we believe research begins before the first bit of data is collected. However, to keep in touch with our user communities, LDC conducts traditional research as well. Recent projects have focused on morphological tagging and syntactic parsing of Arabic, information extraction from biomedical text, the correlation of socioeconomic and linguistic features, the influence of Yoruba on its diaspora and the lexical relationships among Mandé languages.

Knowledge Transfer & Community Service

The institutional knowledge data centers have developed would be of limited use if not shared. LDC transfers knowledge and serves its communities by publishing data, specifications, tools and academic papers, by offering training and by serving on funding panels, program committees and oversight boards.

For more information about how LDC can help your project contact us at: 215-898-0464.

Annotation

Speech

- Bandwidth, signal quality, language, speaker
- Time-aligned orthographic transcription
- Story, turn, word segmentation and alignment
- Discourse structure and disfluency labeling
- Sociolinguistic variables

Translation

- Multiple translation and quality assessment
- Post-editing, MT output evaluation
- Document, sentence, phrase, word level parallel text alignment
- Fine- and coarse-grained topic labeling
- Single and multi- document, monolingual and multilingual summarization

Syntactic and Semantic

- MPG (morphological, part-of-speech, gloss)
- Treebank, PropBank
- Entities, relations, events, time and location
- Co-reference within and across documents
- Entailment
- Knowledge base population
- Sense disambiguation

Image and Video

- Video event, entity labeling and co-reference
- Video and image text transcription
- Image zoning, ground truthing at line, word, sub-word level
- Document classification, legibility, readability
- Human gesture labeling

Lexicons

- Traditional
- Pronunciation with morphology
- Translation

Document

- Topic relevance
- Novel information
- Headlines
- Summarization of varying lengths

Quality Assurance

- Inter-annotator agreement
- Annotation consistency
- Adjudication: human versus human/system
- System output assessment for multiple technologies