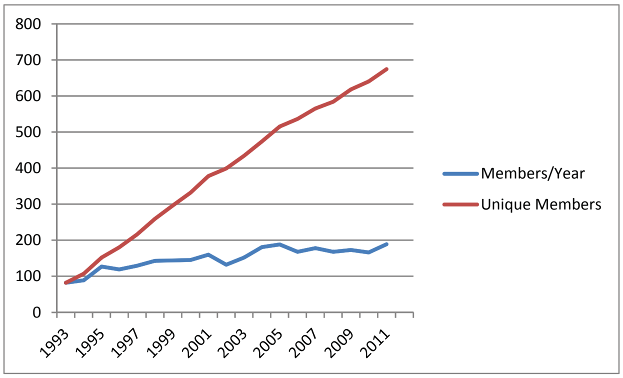
LDC: COLLABORATION

Overview

Collaborations and LDC

Collaborations play an important role in many LDC activities. Through its first two decades, LDC has partnered, consulted and otherwise “collaborated” with a variety of organizations for a multitude of purposes.

Members. LDC would not exist if not for its members who span the range of organizations and fields in language-related research, education and technology development. Universities, research labs, companies and government organizations from around the globe use LDC data for their work. And the mutual interest underlying the member relationship is often the seed that spawns additional joint undertakings.



A core of approximately 100 members join LDC each year. New members sign up continually.

Data Contributors. LDC’s core mission is to develop, acquire and distribute language-related data, making data contributors valued partners. These include organizations whose business involves data of interest (news organizations, broadcasters) and researchers who contribute corpora and related resources for the catalog.

Sponsors, Research Partners. For sponsored projects, common task evaluations and other specific work, LDC partners with US government agencies (including the Departments of Commerce, Defense, Education, Homeland Security, Interior, Justice, Treasury), the National Science Foundation, research institutions, commercial organizations and others to achieve shared goals. This always results in the creation of language resources that are shared more broadly.

Sister Organizations, Networks. LDC is an active participant in related consortia and groups whose aim is to advance the ways in which resources are developed and distributed. These include initiatives for standardizing specifications and best practices and for developing new architecture for language resource delivery.

United Nations

LDC’s relationship with the United Nations was formed in 1993 with the joint goal of processing and distributing UN parallel text data in support of linguistic research. Further areas of interest include:

- providing the UN with LDC alignment tools and assistance applying voice and speaker recognition software
- providing LDC with audio recordings of public UN sessions for linguistic research
- developing opportunities for training, internships, research projects and lectures

Data Contributors

LDC’s agreements with over 40 data-providing organizations support the development of training and test data for sponsored projects as well as general corpus creation. Those collections are included in many LDC resources, including the multilingual Gigaword series, parallel text collections and broadcast data sets.

Since 2004, external corpus developers have contributed over 40% of the resources in LDC’s catalog. Just a few of the contributors are:

- Microsoft Research India (Indian language part-of-speech tagsets)
- Brandeis University (TimeBank)
- Indiana University (Nationwide Speech Project)
- University of Georgia (Digital Archive of Southern Speech)
- USC Shoah Foundation Institute (MALACH Interviews and Transcripts)
- MITRE Corp. (aligned transcripts, spatial annotations)
- New York Times (New York Times Annotated Corpus)
- US Military Academy at West Point (multilingual speech databases)
- Charles University Prague (multilingual dependency treebanks)

Research Partners Snapshot

Among the institutions and organizations with whom LDC has partnered are the following:

Hong Kong University of Science and Technology (China): Collection and annotation of Chinese conversational telephone speech and broadcast speech

Brno University of Technology (Czech Republic): Collection, annotation and distribution of multilingual speech data, tools and related resources

Budapest University of Technology and Economics (Hungary): Development of Hungarian and Kurdish NLP technologies and resources

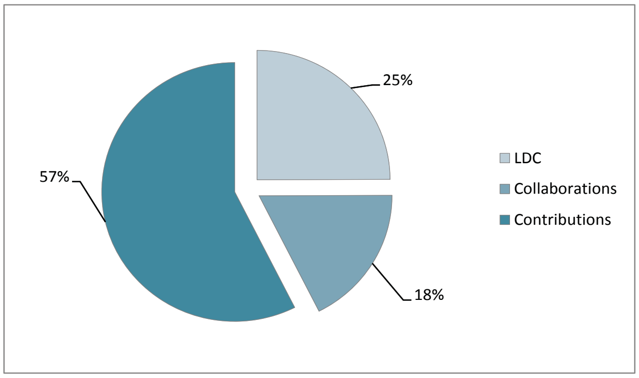
European Language Resources Association/Evaluations and Language resources Distribution Agency (France): Collection and annotation of Arabic broadcast speech

Georgetown University Press (USA): Development of Arabic dialectal dictionaries

Google Inc. (USA): Syntactic structure annotation of English web text

Institut Royal de la Culture Amazighe (Morocco): Development of language resources for Amazigh

Al Akhawayn University (Morocco): Enhancements to LDC's Arabic reading tool and lexical resources development for an Iraqi Arabic WordNet



More than half of LDC's language resources are contributed; the remainder are from collaborations or are developed by LDC.

Vassar College, Brandeis University (USA): Development and deployment of a Language Application Grid to provide access to NLP tools and resources

University of Colorado (USA): Development of multilingual treebanks and propbanks

Columbia University (USA): Development of Arabic tools and language resources

MediaNet (Tunis): Collection of Arabic broadcast speech

Sister Organizations, Networks

LDC works with the European Language Resources Association, the Linguistic Data Consortium for Indian Languages, Gengo-Shigo-Kyokai and others regarding the role of data centers in language resource development and distribution.

LDC collaborates with global networks including the British National Corpus Consortium, E-MELD, European projects such as CLARIN, ENABLER, FLaReNet and META-NET, the Japan-based Language Grid and the US TalkBank project.

LDC is a member of the Open Language Archives Community (OLAC), an international partnership to create a worldwide virtual library of language resource metadata, which includes consensus for best practices for digital archiving. LDC's catalog (searchable through OLAC) consistently receives OLAC's five-star rating for overall metadata quality.

As part of the American National Corpus Consortium, LDC contributed data to this corpus development effort and supports the continued broad availability of the American National Corpus and its progeny through the LDC Catalog and other means.

LDC serves multiple research communities by its representation on funding panels, editorial boards, scientific committees and as conference and workshop participants.

Interested in Collaborating?

LDC welcomes new collaborations. Let us know what interests you and how we can work together. Contact ldc@ldc.upenn.edu to begin the conversation.