



# LDC Activities Related to Languages of the Americas

*Christopher Cieri, Mark Liberman, Denise DiPersio, James Fiumara*

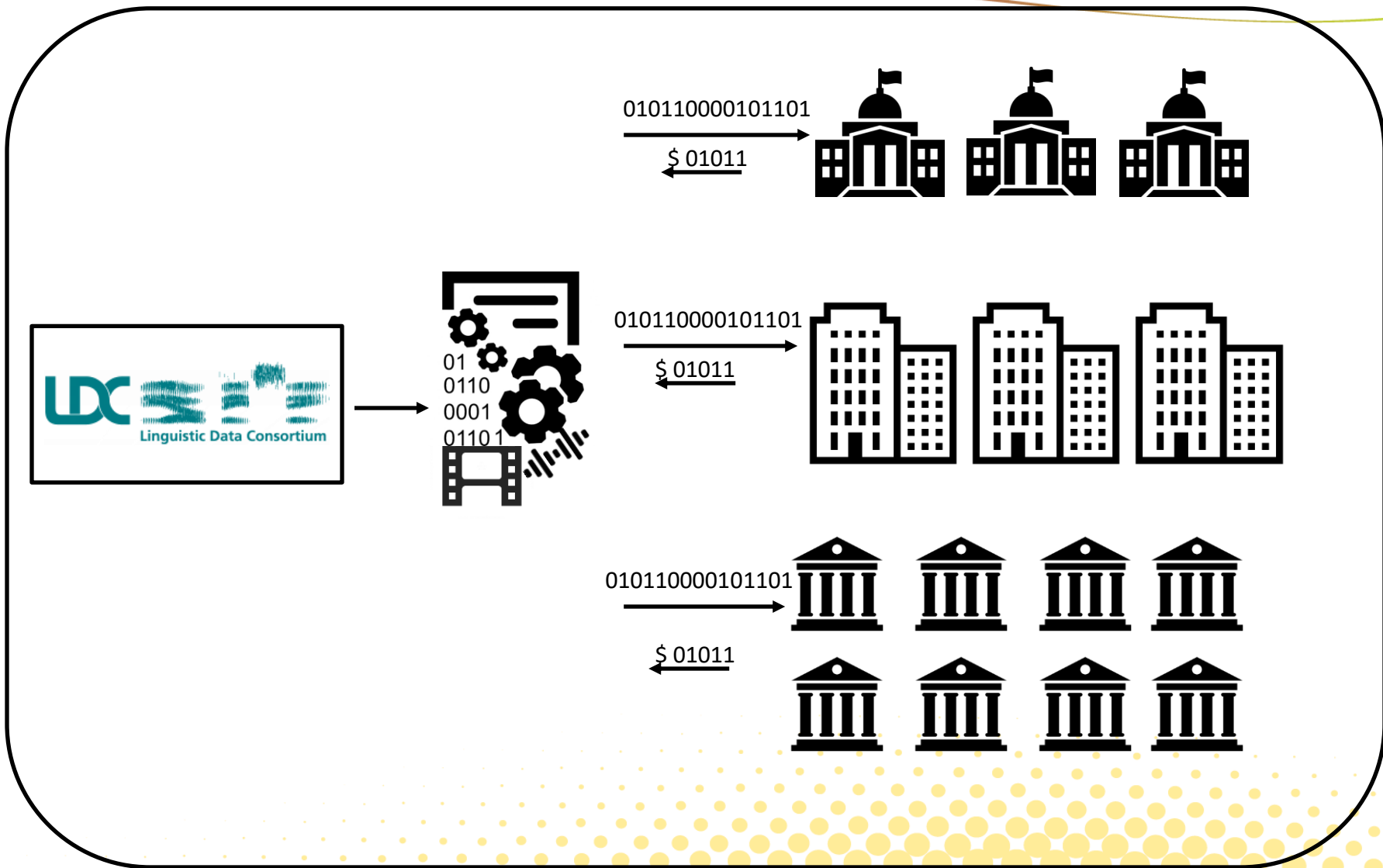
*University of Pennsylvania, Linguistic Data Consortium*

*{ccieri, myl, dipersio, jfiumara} AT ldc.upenn.edu*

- ◆ Support for Networking in US < Europe
  - NetDC, Digging into Data experiences
  - FlareNet
  - CLARIN
- ◆ COCOSDA
  - Early: LTTS Server
  - Current: biennial with regional, topical reports
  - OCOCOSDA: annual, contribution on languages of Asia
  - Shortcomings: speech only, isolates data creators from users
- ◆ Opportunities
  - Increasing participation from ‘the Americas’
  - Penn Global Initiatives
    - China, Latin America, most recently India
- ◆ Can we organize to better collaborate

- ◆ Some participants commented that you were:
  - not a data center
  - not working specifically on languages of the Americas
- ◆ But You Represent
  - Data creators
  - Institutional Archives
  - 'Regional' Data centers
  - Search Providers
  - Networking Services
- ◆ Working either on
  - Languages of the Americas
  - Language in the Americas
- ◆ Effect of that context is unique (or at least different from Europe) in terms
  - linguistic, demographic, economic, legal/regulatory, cultural
- ◆ Frankly most of you represent multiples of these

- ◆ Meet to sketch functions/approaches
- ◆ Identify Major Challenges
- ◆ Can the challenges faced by any one group be met by any other?
- ◆ Identify challenges in common
- ◆ Evaluate according to importance, urgency
- ◆ Remove boundaries: speech vs. text,
- ◆ Do we have enough in common to justify greater networking, collaboration?
- ◆ If so, what would that look like?



## Sponsors

- contribute project funding
- receive
- needs assessment
- custom LR creation & coordination
- stability (since 1992)
- reliability (IPR, IRB)
- infrastructure
- cost sharing
- innovation

## Authors

- contribute data sets
- receive
- QA, repair **before** publication
- broad exposure: newsletter >8000 researchers
- discoverability
- name recognition



## Members

- contribute member fees
- receive
- ongoing rights
- volume, quality, variety
- high ROI

## Media Providers

- receive
- peace of mind
- efficiency
- consistency

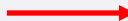
Membership Types	# Corpora	Delivery	Perpetual Rights	Commercial Rights
Standard	16	on demand	✓	
Subscription	ALL (≥36)	automatic	✓	
Commercial Standard	16	on demand	✓	✓
Commercial Subscription	ALL (≥36)	automatic	✓	✓




- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- DATA MANAGEMENT
- COLLABORATIONS

**Quick Links**

- CATALOG
- NEW CORPORA
- USER LOGIN
- HOW TO GET DATA
- DATA MANAGEMENT PLANS
- PROJECTS



**What's New:**

LDC at LREC 2018: Thursday May 10

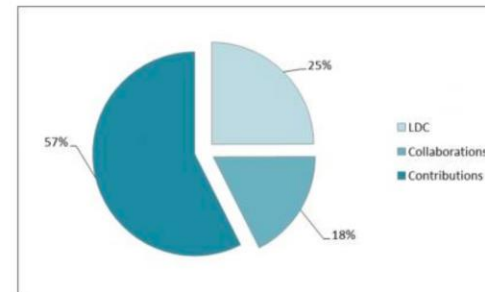
Web pages feature DMPs

LDC enhances its user services

Staff Podcasts Accessible on the LDC Blog

**Interested in Collaborating?**

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



LDC welcomes new collaborations; let us know what interests you and how we can work together. Contact [the External Relations Group](#) to begin the conversation.





[My Account](#) [Logout](#) Bin: (Empty)

- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES ▾
- Data ▾
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora
- Projects
- Search
- Memberships
- LDC Online
- Data Scholarships
- Tools >
- Papers >
- LR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

[Home](#) > [Language Resources](#) > [Data](#)

**Logged in successfully**

## My Account

### Linguistic Data Consortium

**My Address:** ([Edit](#))  
 Christopher Cieri  
 3600 Market St., Suite 810  
 Philadelphia, Pennsylvania  
 19104  
 United States

[ccieri@ldc.upenn.edu](mailto:ccieri@ldc.upenn.edu)  
 ([Edit e-mail](#) or [password](#))

**Organization Contact:**  
 Caitlin Fontecchio  
 3600 Market St  
 Suite 810  
 Philadelphia, Pennsylvania  
 19104  
 United States  
 215-573-1275  
[caitlifo@ldc.upenn.edu](mailto:caitlifo@ldc.upenn.edu)

- **Account Options** —
- [Corpora Invoiced](#)
  - [Agreements](#)
  - [Downloads](#) ←
  - [LDC Online](#)
  - [Receive Newsletter](#)

[LDC Intranet Login](#)  
[LDC WebMail](#)

### — Membership Years —

- 1993 (In Kind, Subscription)
- 1994 (In Kind, Subscription)
- 1995 (In Kind, Subscription)
- 1996 (In Kind, Subscription)












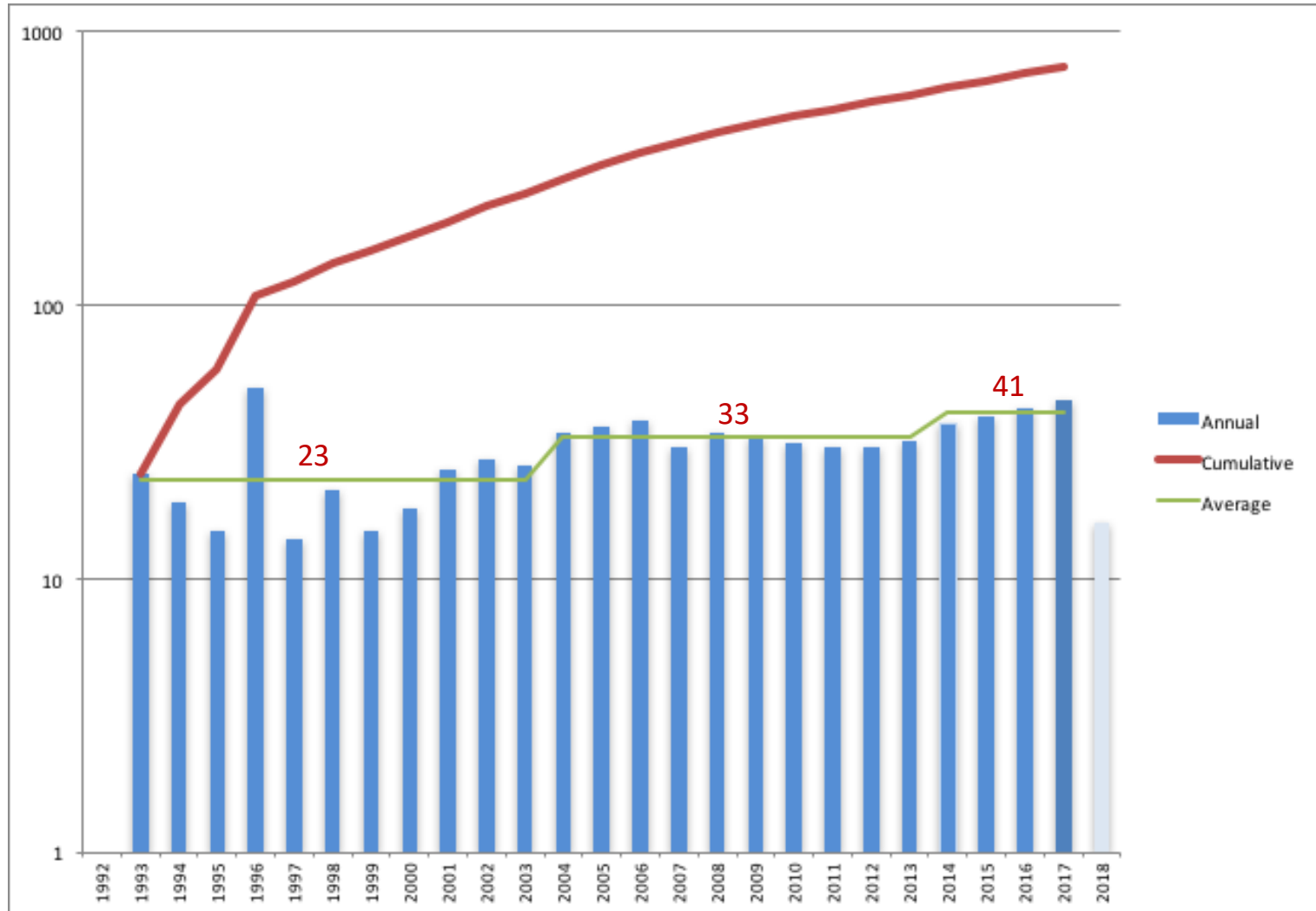
[My Account](#) [Logout](#) Bin: (Empty)

- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES ▾
- Data ▾
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora
- Projects
- Search
- Memberships
- LDC Online
- Data Scholarships
- Tools >
- Papers >
- LR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

[Home](#) > [Language Resources](#) > [Data](#)

## Corpora Available for Download

LDC Catalog ID	Corpus Name	Invoice Date	Download	
LDC2018T09	SPADE	2018-03-15	(0) 	SPADE_LDC2018T09 File Size: 1020 KB MD5 Checksum: 9d2
LDC2018T10	BOLT Arabic Discussion Forums	2018-03-15	(0) 	bolt_ara_disc_for_src_data_LDC2018T File Size: 16 GB MD5 Checksum: 574e
LDC2018T10	BOLT Arabic Discussion Forums	2018-03-15	(0) 	bolt_ara_disc_for_src_data_LDC2018T File Size: 13.8 GB MD5 Checksum: c71
LDC2018T11	LORELEI Somali Representative Language Pack - Monolingual and Parallel Text	2018-03-15	(0) 	lorelei_somali_mono_para_txt_LDC201 File Size: 293 MB MD5 Checksum: 7ac
LDC2018S02	IARPA Babel Tok Pisin Language Pack IARPA-babel207b-v1.0e	2018-02-16	(0) 	IARPA_BABEL_OP2_207_LDC2018S0 File Size: 7.6 GB MD5 Checksum: 28d2
LDC2018S03	Multi-Language Conversational Telephone Speech 2011 -- Central Asian	2018-02-16	(0) 	multilang_cts_2011--central_asian_LDC File Size: 1.63 GB MD5 Checksum: 357
LDC2018T04	LORELEI Amharic Representative Language Pack - Monolingual and Parallel Text	2018-02-16	(0) 	lorelei_amharic_lang_pack_mono_para File Size: 570 MB MD5 Checksum: b95
LDC2018T03	TAC KBP Comprehensive English Source Corpora 2009-2014	2018-02-16	(0) 	tac_kbp_comp_eng_src_2009 File Size: 6.54 GB MD5 Checksum: e2f
LDC2018T07	2007 CoNLL Shared Task - Greek, Hungarian & Italian	2018-01-26	(0) 	conll_2007_gre_hung_ita_LDC2018T07 File Size: 3.73 MB MD5 Checksum: 97a
				conll_2007_ara_eng_LDC2018T08



- ◆ 169 data sets since last report, 2014
- ◆ 55 corpora from DARPA GALE
  - multiple Arabic & Chinese broadcast news and broadcast conversation with transcripts
  - Arabic-English and Chinese-English parallel text from
    - broadcast news
    - broadcast conversation transcripts
    - newswire
    - web text
  - word-level alignments of the parallel text
  - Arabic-English & Chinese-English parallel aligned Treebanks
- ◆ 15 IARPA Babel languages packs
  - Assamese, Bengali, Cantonese, Georgian, Haitian, Kurdish, Lao, Pashto, Swahili, Tagalog, Tamil, Tok Pisin, Turkish, Vietnamese and Zulu
- ◆ The first of the LORELEI language packs in Amharic & Somali

- ◆ 64 Spanish releases (as of May 2018)
  - Broadcast audio and video: news and conversation
    - Broadcast speech and transcripts: HUB4 evaluation, CIEMPIESS, CIEMPIESS LIGHT
  - Conversational telephone speech
    - CALLHOME, CALLFRIEND (Caribbean, non-Caribbean dialects), HUB5 evaluation, Fisher, Fisher and CALLHOME Transcripts English translation, VAHA (Polyphone II), Syllable-Final /s/ Lenition
  - Microphone speech
    - LATINO-40 Spanish Read News, Hispanic-English Database (native speakers from Central and South America) West Point Heroico Spanish Speech (portions recorded at Heroico military academy in Mexico City), United Nations Proceedings Speech
  - Newswire, laws, government documents, web text
    - Spanish Gigaword (3<sup>rd</sup> Edition), United Nations parallel text, ECI Multilingual Text, NewSoMe Corpora of Opinions in Blogs and News Reports, TREC Spanish

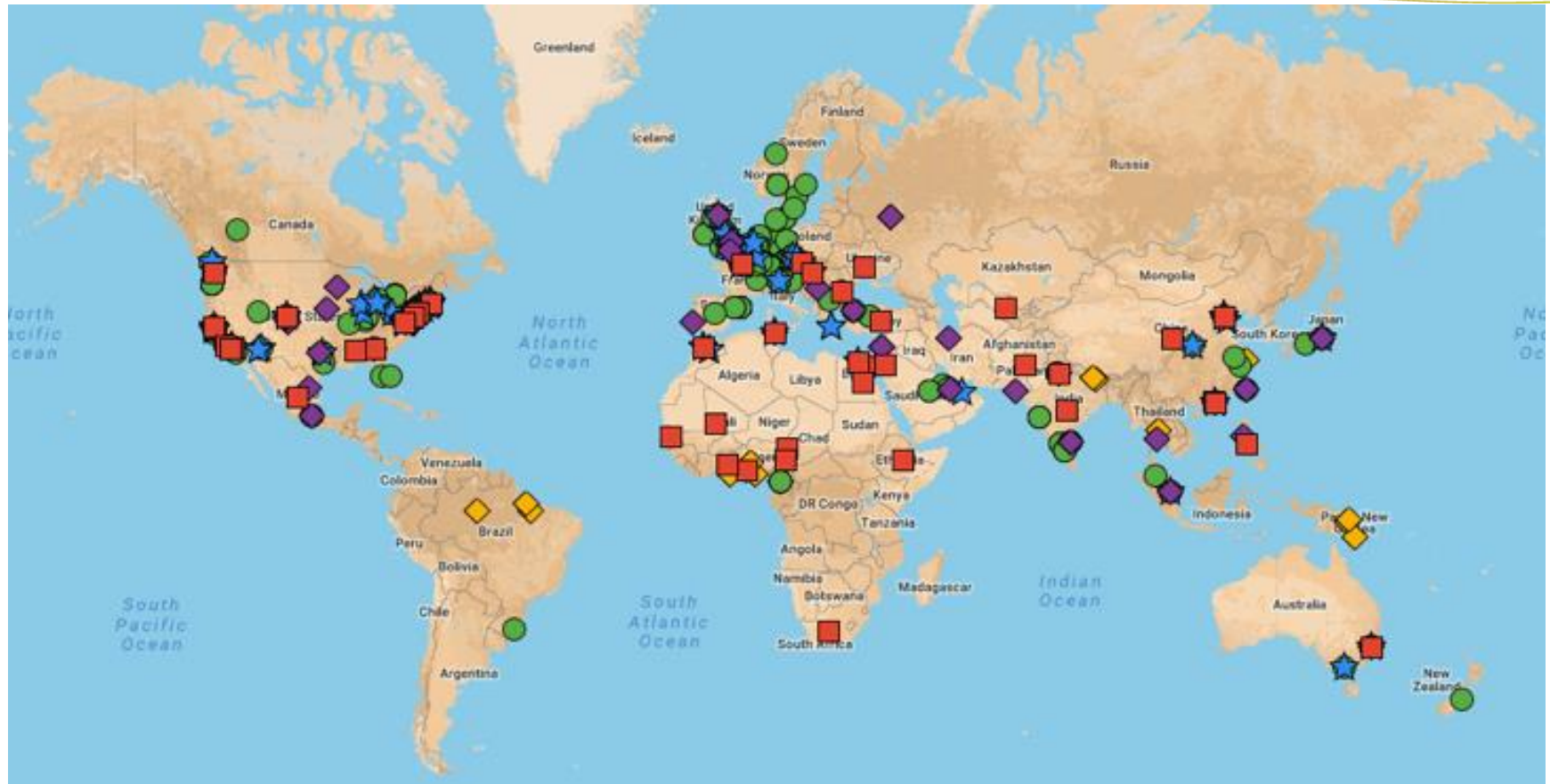
- Syntactic and semantic annotation
  - DEFT Spanish Treebank (Penn Treebank phrase structure), Spanish dependency treebanks (CoNNL shared tasks)
  - ACE 2007 (entities, information extraction, time normalization)
  - ModeS TimeBank, Spanish TimeBank (temporal, event, spatial annotation)
  - SenSem Lexicon (verb feature descriptions); SenSem Databank (semantic annotation)
  - Event/Status Corpus (civil unrest events with temporal tags)
  - TAC KBP Spanish Cross-lingual Entity Linking (queries, entity information, knowledge base links)
  - CALLHOME Spanish Dialogue Act Annotation (discourse)
- Donations
  - National Autonomous University of Mexico (UNAM)
  - Technical University of Madrid
  - Barcelona Media
  - GRIAL Inter-University Research Group
  - More donations welcome!

- ◆ The LDC Catalog has served as a permanent repository for language resources since its inception
  - Seeded first by data contributions of significant corpora – TIMIT, TIDIGITS, CSR
  - Augmented over time by data developed by LDC in sponsored projects: ACE, ACQUIANT, BOLT, DEFT, EARS, GALE, HAVIC, MADCAT, TDT and the NIST evaluation series (LRE, SRE, OpenMT, TREC, MED, MER, TAC KBP) along with contributions from the global research community
- ◆ Metadata and catalog descriptions adhere to established information science standards and digital repository best practices
  - Consistent 5-star rating for metadata quality awarded by OLAC
- ◆ LDC Catalog certified as a CoreTrustSeal Data Repository in 2018
  - The Catalog meets a series of requirements covering data access, rights management, curation, storage and so on developed by the ISCU World Data System and the Data Seal of Approval
  - LDC joins 130+ other data repositories in their commitment to promoting sustainable and trustworthy data infrastructures

- ◆ 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 150,000 copies distributed under various license arrangements
  
- ◆ Non-exclusive rights to LDC
  - Consistent with mission to provide broad access to data



- ◆ Reissuing legacy corpora
  - Updating encoding, formats, metadata, documentation
  
- ◆ Catalog and business system enhancements
  - Incorporates e-commerce principles
  - Users have increased control over their LDC accounts
  - Can license data and join LDC online
  
- ◆ 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 Global Network of select data sources including subcontractors and vendors (red squares), corpus authors (green circles), media providers (purple diamonds), LDC staff collections (gold diamonds), research collaborators (blue stars). Many markers represent multiple collaborators; many markers partially obscured by others.

- ◆ text from news sources, journals, financial, biomedical documents
- ◆ internet sources including newsgroups, (micro)blogs and discussion fora
- ◆ text interactions via email, chat and SMS
- ◆ printed, handwritten and hybrid documents, for example printed forms completed by hand
- ◆ audiovisual data from broadcast news and conversation, podcasts, conversational telephone speech, lectures, interviews, meetings, field interviews, read and prompted speech, task oriented speech, role play, speech in noise, web video and even animal vocalizations
- ◆ digitized analog media including interviews in a variety of tape formats.

- ◆ data scouting, data triage and smart data selection
- ◆ alignment of paired audio, auditing of bandwidth, signal quality, language, dialect, program, speaker
- ◆ quick, quick-rich and careful transcription, audio segmentation and audio-text alignment at story, turn, sentence, word level
- ◆ orthographic, spelling and phonetic script normalization and transliteration
- ◆ tagging of phonetic, dialect, sociolinguistic and supralexic features
- ◆ document zoning, handwriting transcription, OCR QC and tagging of reading order
- ◆ tokenization, tagging of morphology, part-of-speech and gloss, Treebanking, PropBanking, SemBanking
- ◆ sense disambiguation, fine and coarse-grained topic relevance annotation
- ◆ novelty, textual entailment, hypothesis generation and inference annotation
- ◆ annotation of committed belief, sentiment, disfluency, discourse features and hedging
- ◆ detection and classification of entities, relations, events, time, location and their co-reference in text, knowledge base population
- ◆ single and multi-document summarization of various lengths from titles to 200 words
- ◆ query generation and question answering
- ◆ translation, multiple translation, edit distance, translation post-editing and quality control
- ◆ alignment of translated text at document, sentence, phrase & word levels
- ◆ describing the physics of gesture via joint angles and rotations
- ◆ identification, classification and tracking entities and events in video
- ◆ assessment of IR, MT, KBP, QA and other system output

- ◆ DEFT: Deep Exploration and Filtering of Text Program
  - “address remaining capability gaps related to inference, causal relationships and anomaly detection”
  - IE about events and Sentiment/ Emotive/ Cognitive state (SEC), in addition to entities and relations.
  - LDC annotated news text and discussion fora for Entities, Relations and Events (ERE), Abstract Meaning Representation (AMR), Textual Entailment and Committed Belief
- ◆ Conflicting Accounts of Current Events (CACE)
  - multimedia of multiple accounts in varying formality of current events (news, blogs, discussion forums, microblogs video, image, speech)
  - multiple topics per scenario; topic model specifying salient entities & events
  - subset labeled for entities, events, arguments, attributes (i.e. slots), lightweight SEC (Sentiment/ Emotion/ Cognitive State), entailment & contradiction relationships
  - continuing in AIDA
- ◆ Low Resource Languages for Emergent Incidents (LORELEI)
  - situational awareness (topics, names, events, sentiment, relationships) in LRL sources via Machine Translation, Named Entity Recognition, Entity Discovery and Linking, Situation Frame
  - LDC: text NE, coref, NP chunking, lightweight SEC, EL, situation frames, lexicon, grammatical sketch, tokenizers, encoding converters, segmenters and entity tagger
  - representative languages: Hausa, Turkish, Amharic, Arabic, Somali, Farsi, Russian, Spanish, Hungarian, Mandarin, Vietnamese, Yoruba, Tamil, Bengali, Hindi, Indonesian, Tagalog, Thai, Akan (Twi), Swahili, Wolof, Zulu and Uzbek
  - emergent language: Uyghur, Tigrinya and Oromo.

- ◆ Autism Spectrum Disorders
  - with CHOP CAR
  - conversation in diagnostic sessions with gold-standard diagnoses
  - classifiers based on linguistic features predict diagnoses
  - identified new sex-linked variation (Parish-Morris et al., 2017)
  - preparing to repeat with normal conversations
- ◆ Neuro- Degenerative Disorders
  - HUP Fronto-Temporal Disorder laboratory
  - semi-structured interactions with bvFTD, matched healthy controls
  - fundamental frequency and log-scale pitch range controlling for individual & sex differences, correlated to neuropsychiatric tests, gray matter atrophy
  - bvFTD patients significantly reduced  $f_0$  vs. healthy controls reflecting impaired prosody, supporting feasibility of automated analysis (Nevler et al., 2017)
- ◆ Framingham Heart Study
  - just beginning

- ◆ Is diarization a solved problem?
- ◆ JSALT 2017 asked how systems that perform well on CTS generalize to
  - child language
  - clinical interviews
  - reverberant environments
  - noisy public environment
  - web video
- ◆ DiHard
  - Baidu, Laboratoire de Sciences Cognitives et Psycholinguistique, University of Science and Technology of China, Indian Institute of Science, LDC
  - LDC contributed data, selection, annotation, QC, scoring, logistics
  - to be reported out at Interspeech
  - Want a Hint?

## ◆ NIEUW

- NSF supported project to increase scale of data development
- augmenting current methods with novel (non-monetary) incentives
- developing toolkit to implement games and citizen linguist activities

## ◆ Disaster Recovery Plan

- LDC data replicated remotely via cloud services
- in addition to protections already in place at LDC

## ◆ Media Data ReIndexing, iRODS implementation underway

## ◆ Core Trust Seal

- ICSU World Data System (WDS) & Data Seal of Approval (DSA) merged certification under Research Data Alliance umbrella
- certification based on DSA-WDS Core Trustworthy Data Repositories Requirements catalogue and procedures
- 28 data centers (worldwide, all fields) currently hold this certification