A Progress Report on Activities at the Linguistic Data Consortium Benefitting the LREC Community

Christopher Cieri, James Fiumara, Stephanie Strassel,
Jon Wright, Denise DiPersio, Mark Liberman
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Philadelphia, PA 19104 USA
{ccieri, jfiumara, strassel, jdwright, dipersio, myl} AT ldc.upenn.edu

Abstract

This latest in a series of Linguistic Data Consortium (LDC) progress reports to the LREC community does not describe any single language resource, evaluation campaign or technology but sketches the activities, since the last report, of a data center devoted to supporting the work of LREC attendees among other research communities. Specifically, we describe 96 new corpora released in 2018-2020 to date, a new technology evaluation campaign, ongoing activities to support multiple common task human language technology programs, and innovations to advance the methodology of language data collection and annotation.

Keywords: language resources, common task programs, technology evaluations, citizen linguists

1. Introduction

Linguistic Data Consortium (LDC) activities include the collection, annotation, processing, distribution, archiving and curation of language resources (LRs) including data, tools and specifications to support language related research, education and technology development worldwide and in any language. LDC also conducts data centered research in a growing number of fields that rely upon language data and engages with the relevant research communities to establish best practices in the application of linguistic analysis and human language technologies to their disciplines.

2. The Consortium Model

In 1992 the first goal of the newly formed LDC was to streamline the distribution of LRs, principally data sets. The consortium model developed by the original LDC Advisory Board in 1992 has proven effective and withstood the test of time. Notwithstanding the other successful models of ELRA/ELDA [1], CLARIN [2] and SADiLaR[3], the LDC model persists because it balances between two inevitable but conflicting facts of life: 1) data wants to be free and 2) data creation has its costs and creators need support.

Some motivations behind the free data movement surrounding LRs are that broad access to data: encourages the development and spread of technologies that, for example, raise the standard of living; improves the free flow of other sources of information; reinforces the basic human right to use one's native language; and lowers barriers to participation in the worldwide digital economy thereby reducing the gap between linguistic haves and have-nots.

Nevertheless, there are inevitable costs associated with the creation, distribution and maintenance of LRs. These include not only technical costs such as computing cycles, storage, data security and networking bandwidth, but fees and associated legal costs for licensing copyrighted materials, compensation for human subjects and the researchers who manage corpus development and distribution, and the cost of the facilities where they work.

LDC's principal model attempts to find a stable balance among these competing forces by pooling small contributions from many member organizations to support the people who: provide LRs to members without additional costs; assure that the LRs remain available for use as benchmark data and compete for the chance to create new LRs. For an annual cost less than that of a high end laptop or travel to an international conference, an entire University or research organization gains rights to use many datasets for which individual creation costs are one to three orders of magnitude higher.

Membership in the Consortium is generally the most cost-effective way to access data. However, to accommodate some users, LDC also offers licenses to most individual corpora. Figure 1 shows the locations of members and licensees around the world. Each map point represents 1 to 101 organizations at that location. Particularly encouraging are the large number of new participants, especially in South America, Africa and Southeast Asia.

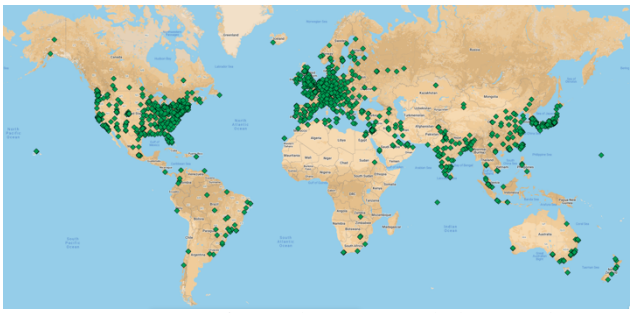

Figure 1: Locations of LDC data users including members and licensees. Each map point represents a city which may have between 1 and 101 organizations and individual users.

3. Data Distribution

The LDC Catalog of datasets is a shared resource created by and for the research communities that LDC serves. More than half of the datasets are contributed by member organizations; the remainder result either from a direct collaboration between LDC and other researchers or from

[1] http://www.elra.info
[2] https://www.clarin.eu
[3] https://www.sadilar.org

common task technology development programs in which LDC has competed and been selected to create resources to support program specific goals.

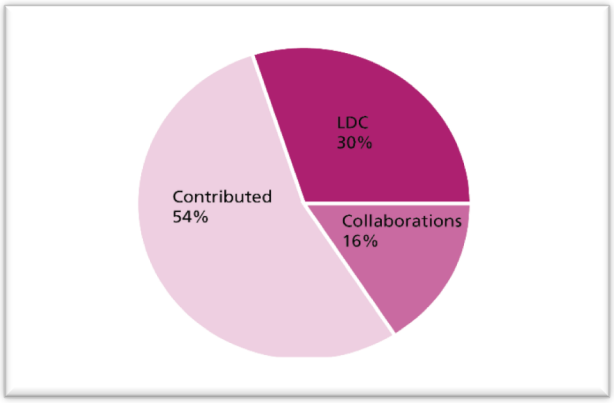

Figure 2: LDC publications by source

Each month LDC releases at least three data sets, of different types to meet a wide range of research needs including: large text corpora, transcribed broadcast news, conversational telephone speech, parallel text, lexicons and treebanks. In 2018 through March 2020, LDC has released 96 corpora. As Figure 3 shows, the number of corpora released per year has been higher and more stable over the past two decades compared to the first. The 10-year average of publications per year, marked by the green line, has grown from 23 to 33 to 42 so far in the current decade. The total number of corpora released, indicated by the red line, is 841. Below, we document the corpora released in 2018-2020 to demonstrate the range of research areas supported and provide LDC Catalog IDs in parentheses for those who wish to further explore specific corpora[4].

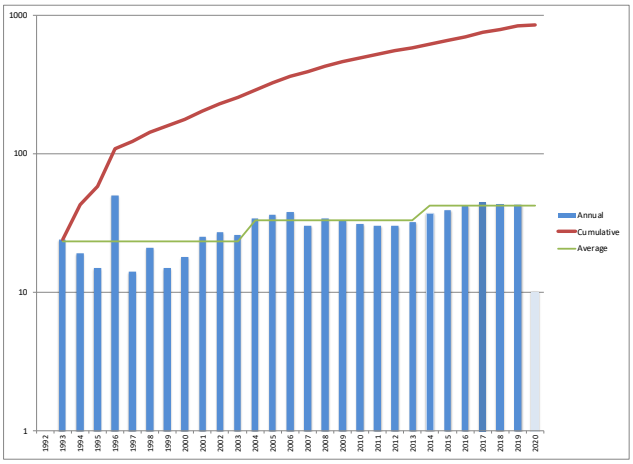

Figure 3: LDC publication rates showing the number of data sets per year in blue bars, 10-year average on the green line and cumulative total on the red line with values on a log scale y-axis.

3.1 Language Recognition

Language Recognition corpora typically contain speech collected via one or more channels of: telephone conversation, broadcast narrowband speech (broadcasts of telephone calls) or broadcast news, typically without transcripts. Such corpora often strive to avoid the situation in which one may predict language on the basis of channel or vice-versa. For example they may record multiple languages within one country or one language in multiple countries or both. From 2018 through March 2020, LDC has released:

- Second editions of the *CALLFRIEND* corpora in American English-Non-Southern Dialect (LDC2019S21), Canadian French (LDC2019S18), Egyptian Arabic (LDC2019S04) and Mandarin Chinese-Mainland Dialect (LDC2018S09) updated by LDC to provide the audio in wav format, improve documentation and simplify directory structure
- 2011 *NIST Language Recognition Evaluation Test Set* (LDC2018S06)
- *Multi-Language Conversational Telephone Speech* collected by LDC in Iraqi, Levantine, Maghreb Arabic (LDC2019S02), Czech, Slovak (LDC2018S08), Dari, Farsi, Pashto (LDC2018S03), American and South Asian English (LDC2019S06), Thai, Lao (LDC2019S15) and Spanish (LDC2018S12)

3.2 Speaker Recognition and Diarization

Corpora that support the development and evaluation of Speaker Recognition technologies typically contain multiple speech samples from each of many speakers recorded over a time span of days to weeks. If channel is varied during the collection, there is typically an effort to assure that many speakers are recorded over the same or very similar channels. Much of the speaker recognition data LDC has released has come from the Mixer initiative, an effort to systematically covary channel, language and communicative situation over large collections of speakers in the US. Publications include:

- *Mixer 4 and 5* (LDC2020S03) telephone conversations, in-person interviews, read and prompted speech from 616 speakers that was used in the 2008 NIST Speaker Recognition Evaluation.
- NIST 2016 Speaker Recognition Evaluation Test Set (LDC2019S20)
- Development (LDC2019S09, LDC2019S10) and evaluation (LDC2019S12, LDC2019S13) data used in the first *DIHARD Challenge*
- 29 hours of Mandarin Chinese audio from user contributed video with transcripts (LDC2019S05) developed by LDC for the *Video Annotation for Speech Technologies* (VAST) project, to support speech activity detection, language identification, and speech recognition as well as speaker identification from a novel source with a wide range of speakers, domains, and noises.

3.3 Speaker Characterization

Given the lone entry here and generally small number of speaker characterization corpora, it is premature to try to list features that distinguish this category.

- *Nautilus Speaker Characterization* corpus, contributed by the Technical University of Berlin, containing 155 hours of conversational German speech from 300 speakers produced during scripted and semi-spontaneous dialogs including

[4] Descriptions of all LDC corpora are permanently available by adding the CatalogID to: https://catalog.ldc.upenn.edu/

spontaneous neutral and emotional speech utterances, evaluated for perceived interpersonal speaker characteristics (e.g. likable, attractive, competent, childish) and a subset for naive voice descriptions (e.g. bright, creaky, articulate, melodious)

3.4 Speech Recognition and Synthesis

Corpora designed for Speech Recognition and Synthesis typically contain speech and associated transcripts, often time-aligned at the utterance, word or phone level. Speech recognition corpora often target real world recording conditions while corpora for speech synthesis typically strive for the highest possible sound quality. LDC releases in these two categories include:

- Language packs for Amharic (LDC2019S22), Cebuano (LDC2018S07), Dholuo (LDC2020S02), Guarani (LDC2019S08), Igbo (LDC2019S16), Kazakh (LDC2018S13), Lithuanian (LDC2019S03), Telugu (LDC2018S16) and Tok Pisin (LDC2018S02) all collected by Appen[5] and processed by the *IARPA Babel* program for use by participants.
- *Arabic Broadcast News Speech* (LDC2018S05) and *Transcripts* (LDC2018T14) created for the DARPA GALE program, Phase 4
- A second edition of *Mandarin Telephone Speech and Transcripts* (LDC2018S18)
- *Magic Data Chinese Mandarin Conversational Speech* containing 10 hours of transcribed from 60 speakers, recorded on multiple devices (LDC2019S23) and contributed by Beijing Magic Data
- *AISHELL-1* (LDC2018S14), contributed by Beijing Shell Shell Technology, containing ~520 hours of Mandarin readings of sentences representing 11 genres from 400 speakers from the North, South and Yue-Gui-Min regions of China
- *Polish Speech Database* (LDC2019S19), from VoiceLab [6] containing 263,424 transcribed utterances, ~280 hours in total, from 200 speakers who recorded themselves for at least 60 minutes from their home computers using a headset and reading text from a website
- *USC-SFI MALACH Interviews and Transcripts English, Speech Recognition Ed*. (LDC2019S11) containing ~168 hours of interviews from 682 Holocaust witnesses along with transcripts, a lexicon, and Kaldi specific files
- *DIRHA English WSJ Audio* (LDC2018S01), developed by the Distant-Speech Interaction for Robust Home Applications Project[7], containing ~85 hours of real and simulated read speech by six native American English speakers collected in an apartment with typical background noise and reverberation via 32 microphones distributed around the space with annotations for microphone positions, speaker id, gender and position
- two corpora contributed by the Development of Speech Technologies program at the National Autonomous University of Mexico: *CIEMPIESS Balance* (LDC2018S11) containing 18 hours of Mexican Spanish broadcast speech with transcripts and *CIEMPIESS Experimentation* (LDC2019S07) containing 22 hours of transcribed Mexican Spanish broadcast and read speech including a phonetically-balanced set of isolated words, recordings of 21 females speakers to create gender balance with other CIEMPIESS collections, and 10 hours of new test data
- *LibriVox Spanish* (LDC2020S01), contributed by Carlos Daniel Hernández Mena and his students, consisting of ~73 hours of audiobook reading developed by the LibriVox [8] project with transcripts provided by native Spanish speakers
- *Avatar Education Portuguese* (LDC2018S15) containing 1,400 utterances of read and spontaneous Brazilian Portuguese, transcribed at the word and phoneme levels and contributed by the University of Pernambuco

3.5 (Multimedia) Information Retrieval

Corpora to support Information Retrieval are traditionally built from news text but increasing include web text, such as discussion forums, or other media, such as user contributed videos. These corpora also typically contain annotations of relevance to event, topic or specific questions. Recent publications of this type are:

- *HAVIC MED Progress Test - Videos, Metadata and Annotation* (LDC2018V01, LDC2019V01) developed by LDC and NIST comprising ~3,650 hours of user-created videos selected and annotated either to represent defined event types or to serve as distractors and labelled with topic and genre to support the NIST Multimedia Event Detection task and to measure progress over time
- *BOLT Information Retrieval Comprehensive Training and Evaluation* (LDC2018T18) including discussion forums in Arabic, Chinese, and English, natural-language queries, system responses, human assessments and scoring software, developed by LDC to support the DARPA BOLT program (see §4.4)

3.6 Information Extraction

Corpora for Information Extraction are commonly built from text, though the types have been diversifying over time. Annotation often includes entities, events, relations and coreference and may be normalized or linked within or across documents, sometimes across languages and increasingly to knowledgebases.

- *Committed Belief Annotation* of discussion forums in Chinese (LDC2019T03), English (LDC2019T16) and Spanish (LDC2019T09) developed by LDC for the DARPA DEFT program described in §4.5
- A collection (LDC2019T14) of 110 source documents from English newswire manually annotated for instances of categories defined with respect to the NFL Scoring ontology (Strassel et al. 2010) to support the DARPA Machine Reading program

[5] https://appen.com
[6] https://www.voicelab.ai
[7] https://dirha.fbk.eu
[8] https://librivox.org

- Multiple datasets from the NIST TAC KBP task including: *source* data for 2009-2014 (LDC2018T03) and 2016-2017 (LDC2019T12); training and evaluation data for the *Slot Filling* task in English for 2009-2014 (LDC2018T22) and in Chinese for 2014 (LDC2019T08); evaluation data for the 2012-2017 *Cold Start* task (LDC2019T17); training and evaluation data for the *Entity Discovery and Linking* task for 2009-2013 (LDC2018T16), 2014-2015, (LDC2019T02) and 2016-2017 (LDC2019T19), the *Relation Extraction* task (LDC2018T24) and the *English Event Argument* task 2014-2015 (LDC2020T03)
- *Machine Reading P1 IC Training* (LDC2020T04) developed by LDC for the DARPA program which sought to extract knowledge from natural language text for use in formal reasoning systems in multiple domains. The corpus contains 248 English newswire documents (~109K words) about half of which are annotated for entities and specific relations (Attack, Biographical, Affiliation and Family), then aligned with an ontology to allow automated reasoning. Annotation was not exhaustive but sought to cover all relations and arguments mentioned explicitly and some that annotators only "inferred".

3.7 Natural Language Processing

The boundaries of this category are not so well defined. Here, we include corpora that support syntactic, semantic and discourse parsing and tagging.

- 2007 *CoNLL Shared Task* data in - Arabic & English (LDC2018T08), Basque, Catalan, Czech & Turkish (LDC2018T06) and Greek, Hungarian & Italian (LDC2018T07)
- *Spanish Treebank* (LDC2018T01) created for the DEFT program (§4.5).
- *Phrase Detectives, Version 2* (LDC2019T10) contributed by the University of Essex which include annotations of discourse new and anaphoric relations acquired via a game of the same name
- The third release of the *Penn Discourse Treebank* (LDC2019T05) containing >53,500 tokens from the Penn Treebank annotated with discourse relations with standardizations, new senses and consistency checks applied to earlier versions
- Two corpora, *Concretely Annotated New York Times* (LDC2018T12) and *Concretely Annotated English Gigaword* (LDC2018T20), contributed by Johns Hopkins University's Human Language Technology Center of Excellence containing automatically generated sentence segmentations and word tokens, constituent parse trees, dependency trees, named entities, part of speech tags, lemmas, entity coreference chains and frame semantic parses of two large news text corpora previously released by LDC
- *Abstract Meaning Representation* (AMR) corpora in both Chinese (LDC2019T07) and English (LDC2020T02). AMR is a syntax-free representation of sentential semantics focused on "*who is doing what to whom*" and built from PropBank frames, semantic roles, entity and coreference annotation, modality, negation, questions and quantities. Chinese AMR was contributed by Brandeis University and Nanjing Normal University and contains representations of 10,325 sentences from the weblog and discussion forum sections of the Chinese Treebank 8.0 using a graph formalism adapted from the English. The English data resulted from a collaboration among LDC, SDL/Language Weaver, University of Colorado and University of Southern California and contains representations of over 59,255 English sentences from newswire, web text, broadcast conversation and fiction.

3.8 Machine Translation

Machine Translation corpora typically contain source language material with translations into the target language that may be found or created for the corpus. Source material may have been born as text or may be transcripts of speech. Source and translation are typically aligned at the sentence level, though word level alignments have proven useful. Many of the 2018-2020 releases are due to the DARPA BOLT program (see §4.4), that focused on (Egyptian) Arabic, (Mandarin) Chinese and English. BOLT releases include:

- Arabic discussion forums (LDC2018T10), translations into English (LDC2019T01), word alignments (LDC2019T06) and a Treebank (LDC2018T23)
- Arabic and Chinese SMS/Chat (LDC2018T15) with English translations (LDC2018T19), word alignment of the Arabic (LDC2019T18) and English translations, word alignments and tagging of the Chinese (LDC2019T13)
- Arabic Conversational Telephone Speech with translations and word-alignments (LDC2020T05)
- Treebank of the English discussion forum text (LDC2019T15)

Other corpora for machine translation include:

- *TRAD Arabic-French* parallel text from newswire (LDC2018T21) and newsgroups (LDC2018T13) and *TRAD Chinese-French* parallel text from broadcast news (LDC2018T17) and blogs (LDC2018T02) created in an LDC/ELRA collaboration
- *SPADE* (LDC2018T09) annotated parse trees and alignment on English sentential paraphrases extracted from machine translation evaluation corpora and augmented with HPSG trees and phrase alignments contributed by Yuki Arase and Junichi Tsujii
- *Multilingual ATIS* (LDC2019T04), from Google, consisting of 5,871 utterances from the ATIS corpora manually translated into Hindi and Turkish, back-translated into English and annotated with named entities such as city, airline, airport names and dates.

3.9 Under-Resourced Languages

This category, also overlapping with others, is marked by corpora for languages with insufficient resources that may support information extraction, translation, natural language processing or other HLTs.

- In addition to the IARPA Babel language packs described in §3.4 above, LDC released two *LORELEI Representative Language Packs*, multi-resource datasets for Amharic (LDC2018T04) and Somali (LDC2018T11) created for the DARPA LORELEI program described in §4.3, the first of many produced by that program.

3.10 Lexicons and Other Datasets

Here we list multiple, inventive lexical datasets and other corpora that do not fit cleanly into the categories above.

- *Database of Word Level Statistics – Mandarin* (LDC2020L01) developed by Hong Kong Polytechnic University, providing romanizations and IPA, syllable structure, length and tone, POS and frequency for Mandarin Chinese words and nonwords
- *EVALution* (LDC2020T06), also developed by Hong Kong Polytechnic University, containing word pairs extracted from English and Chinese Wordnets and ConceptNet 5.0 and annotated by crowd-workers to instantiate semantic relations such as hypernymy, synonymy, antonymy and meronymy and augmented with frequency, part-of-speech.
- *Chinese CogBank* (LDC2020T01), contributed by Bin Li, Siqi Yin, Jie Xu, Li Song and Minxuan Feng contains 232,497 Chinese word-property pairs (雪(snow)-白(white)) collected via Baidu, manually corrected and accompanied by their frequency
- *Rhythm and Pitch* (LDC2018S04), created by Laura C. Dilley, Mara Breen, Meredith Brown and Edward Gibson containing ~27 minutes of English conversations and radio news stories annotated using the 4 tiered Rhythm and Pitch (RaP) scheme to capture intonational and rhythmic aspects of speech
- *Conversational Persian Transcripts* (LDC2019T11) contributed by Ariana Negar Mohammadi, containing transcripts from ~20 hours of informal conversations in Tehrani Persian annotated for gender, age, and recording method and setting
- *DKU-JNU-EMA Electromagnetic Articulography* (LDC2019S14) developed by Duke Kunshan University and Jinan University containing ~10 hours of speech in Mandarin, Cantonese, Hakka, and Teochew Chinese from 2-7 native speakers for each dialect with EMA traces. Speakers read complete sentences, short texts, related words of a specific common consonant, vowel or tone.
- *SRI Speech-Based Collaborative Learning* (LDC2019S01) containing ~120 hours of English speech from 134 US middle school students doing collaborative work, with transcripts and annotations, developed during an SRI initiative to identify patterns in speech that correlate with collaborative learning and to assess collaboration quality
- *H2, E2, ERK1 Children's Writing* (LDC2018T05) contributed developed by the Cooperative State University Baden-Württemberg [9] consisting of ~2,000 stories written in the classroom from picture prompts by 173 German school children, including original and spelling corrected texts, with annotations of names, foreign words, and syntax errors and metadata to indicate school type, age, gender, grade, language spoken

4. Data for Common Task Programs

Much of the data created by LDC supports common task technology development and evaluation programs. To meet the ever growing demand for greater volume, quality and diversity of LRs, LDC has created a global network of contributors who collaborate with the management team in Philadelphia to collect and annotate data in more than one hundred languages and counting. Figure 4 shows the number and diversity of sources including LDC employees and contractors doing new data creation, corpus authors who contribute complete data sets, media organizations that supply source data, and research collaborators who work with us on specifications, technologies evaluation or technical reports. Especially notable are additions since the last report of contributors in South America, Sub-Saharan Africa and Central and Southeast Asia

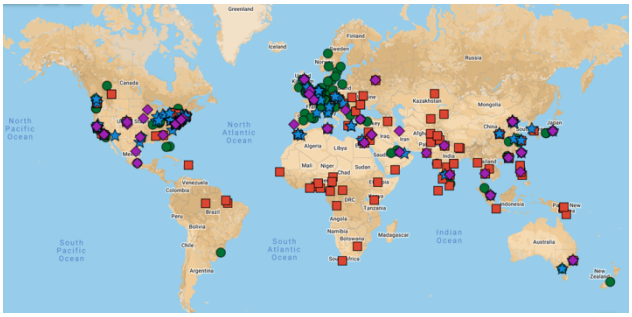

Figure 4: LDC Global Network of data sources: ■ = employees and contractors, ● = corpus authors, ◆ = media providers, ★ = research collaborators. Many locations have multiple contributors; some markers are (partially) obscured.

Even after common task programs end, they continue to pay dividends to multiple research communities through the LRs that appear and, at least in LDC's case, remain available. Below we describe a small sample of the newest and most relevant programs through which LDC is creating and sharing data.

4.1 KAIROS

The official program page describes it goals: "*The Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) … aims to understand complex events described in multimedia inputs by developing a semi-automated system that identifies, links, and temporally sequences their subsidiary elements, the participants involved, as well as the complex event type.*" [10] LDC supports KAIROS by collecting, processing and updating multi-media, multi-lingual data for Schema Learning and Run Time corpora. Schema Learning corpora contain 1,000,000+ documents and 100+ complex events with 5 or more labeled examples each across languages, modalities

[9] http://www.dhbw.de/english/dhbw/about-us.html

[10] https://www.darpa.mil/program/knowledge-directed-artificial-intelligence-reasoning-over-schemas

and sources. Schema Learning corpora are augmented for each evaluation cycle. Run Time data include multiple corpora for development and evaluation with ~5000 new documents across two languages and multiple modalities of which 5-20% are scenario-relevant and 10% are annotated. LDC also assesses system output and distributes KAIROS corpora and related resources. As KAIROS is a new project it has not yet released any corpora for general use.

4.2 AIDA

DARPA describes the goal of its AIDA (Active Interpretation of Disparate Alternatives) program as "*to develop a multihypothesis semantic engine that generates explicit alternative interpretations of events, situations, and trends from a variety of unstructured sources, for use in noisy, conflicting, and potentially deceptive information environments. This engine must be capable of mapping knowledge elements automatically derived from multiple media sources into a common semantic representation, aggregating information derived from those sources, and generating and exploring multiple hypotheses about the events, situations, and trends of interest.*"[11] LDC supports AIDA by collecting multimodal linguistic resources in multiple languages including English Russian and Ukrainian, annotating them for, e.g. topic relevance, entities, events and relations and coreference and assessing system outputs. Although LDC has created numerous pilot data sets for program participants, including multimodal source data in several languages plus pilot annotation for training topic relevance and knowledgebases, AIDA is still in its early phases and has not yet evaluated the data for technology development nor released them into the Catalog.

4.3 LORELEI

DARPA LORELEI seeks to "*dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages. With the understanding that even with perfect translation, there would still be too much material for analysts to use effectively, LORELEI research will not be focused solely on machine translation. While LORELEI technologies may include partial or fully automated speech recognition and/or machine translation, the overall goal will not be translating foreign language material into English but providing situational awareness by identifying elements of information in foreign language and English sources, such as topics, names, events, sentiment and relationships.*"[12] LDC supports LORELEI by creating language packs for each language selected by the program including plain and parallel text, found or newly constructed dictionaries with POS, morphological, entity annotation, situation frames, NP chunking and grammatical sketches. *Representative Language* Packs are intended to provide researchers with data they can use to model the range typological variation present in the worlds languages. LDC has provided LORELEI with representative language packs in the following languages: Akan (Twi), Amharic, Arabic, Bengali, English, Farsi, Hausa, Hindi, Hungarian, Indonesian, Mandarin, Russian, Somali, Spanish, Swahili, Tagalog, Tamil, Thai, Turkish, Uzbek, Vietnamese, Wolof, Yoruba, Zulu. *Incident Language* Packs are intended to simulate the need to quickly build HLTs in a new language to deal with an emergent situation. LDC has provided LORELEI performers with Incident Language Packs in: Uyghur focused on the Xianjiang Earthquake; Oromo and Tigrinya focused on Ethiopia's flood-drought cycle, civil unrest and ethnic violence; and Kinyarwanda/Sinhala focused on Rwanda's drought-flood cycle, civil unrest and refugee crisis and Sri Lankan floods and civil unrest.

4.4 BOLT

The goal of the DARPA BOLT (Broad Operational Language Translation) program was to develop machine translation and information retrieval technologies robust enough to work for informal genres. To support program goals LDC collected text from discussion fora, text messages and chat in Chinese, Egyptian Arabic and English, then translated the non-English data into English with word level alignments and created treebank, Propbank and co-reference annotation. Although the BOLT program has ended, LDC continues to publish corpora the program has funded including 10 since 2018 bringing the total to 18 with more to come. BOLT data contributions to the research community are summarized by language, source and type in Table 1. The linguistic varieties referenced are Egyptian Arabic, Mandarin Chinese and English though regional difference are less apparent in the 'written' varieties. The source types are web text, principally discussion forums (df in the table), and conversation, typically SMS and chat (chat in the table). The data types are the source material, parallel text translated into English, word-alignments of the parallel text and constituency Treebanks.

	Arabic		Chinese		English	
	df	chat	df	chat	df	chat
source	✓	✓	✓	✓	✓	✓
parallel	✓	✓	✓	✓		
word-alignment	✓	✓	✓	✓		
Treebank	✓		✓			

Table 1: BOLT Program contributions to the research community by language, source data and type.

4.5 DEFT

The DARPA DEFT (Deep Exploration and Filtering of Text) program was underway at the time of the last report and has since ended. However the program is relevant in the number of datasets made available to the research community. DEFT "*aims to address remaining capability gaps related to inference, causal relationships and anomaly detection*" [13] by exploring multilingual, multi-document understanding through tradition information augmented by the analysis of sentiment, emotive and cognitive state. LDC provided program participants with news and web text source text that was annotated for ERE (event, entity, relation), AMR (Abstract Meaning Representation), textual entailment and committed belief.

[11] https://www.darpa.mil/program/active-interpretation-of-disparate-alternatives

[12] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

[13] https://www.darpa.mil/program/deep-exploration-and-filtering-of-text

Since 2018, LDC with DEFT sponsorship, has released a Spanish Treebank containing ~110,000 words of news and discussion forums annotated for constituencies plus three corpora of text annotated for committed belief: 950,000 word in English, 83,000 token of Chinese and 67,000 token of Spanish. DEFT has also contributed, with other programs, to the *Abstract Meaning Representation (AMR) Annotation Release 3.0* corpus that was released in January 2020.

5. Technology Evaluation

LDC continues its role of providing data for multiple NIST evaluations including: Speaker Recognition (SRE) in 2018 and 2019 (Sadjadi, et al. 2019), Low Resource Human Language Technologies (LoReHLT) in 2018 and 2019 (Christianson, Duncan, Onyshkevych 2018) and Open Speech Analytic Technologies (OpenSAT) in 2019. LDC also continues to provide data to support community initiatives such as CONLL shared tasks in 2019 (Oepen et al. 2019) and 2020.

In addition, LDC has organized a new evaluation campaign, DIHARD, that seeks to improve diarization robustness to channel, noise and conversational domain. Challenges have taken pace in 2018 and 2019 with results presented at Interspeech workshops specific to the campaign. For the second challenge, data included speech from: audiobooks, broadcast conversation, child language, clinical diagnostic sessions, courtroom debate, map tasks, meetings, restaurant conversations, sociolinguistic field and laboratory interviews, and user contributed web video as well as extracts from the CHiME-5 dinner party corpus (Barker, et al. 2018). DIHARD features four tracks differing by whether the data provided is single or multi-channel and whether the systems do their own speech activity detection or are given the reference SAD. For the single channel condition, there were ~24 hours of development and ~23 hours of evaluation data. For the multichannel condition, systems were provided with ~262.4 hour of development and ~31.24 hours of evaluation data. Human annotation provided the basis for system scoring. The evaluation metrics are DER (Fiscus et al. 2006) and Jacquard Error Rate, developed specifically for DIHARD (Ryant et al. 2019). Performance of a baseline system varied, naturally by track, from a DER of 23.70 on single channel audio using the reference SAD to a high of 87.55 on multichannel audio using the system's own SAD. DIHARD will continue into 2020.

6. Infrastructure Development

LDC's internal web based annotation platform, webann (Wright et al. 2012), continues to support the majority of our workload, adding support for video and image annotation as these media gained importance in LDC projects. With the support of the NSF NIEUW project, we have created a new platform, Universal Annotator (UA), a reboot of the webann code base with portability as an explicit design goal. NIEUW has used UA to build the NameThatLanguage[14] language identification game and the Citizen Linguist portal LanguageARC[15] (see §7) along with the 9 projects currently deployed on that portal and the 5 prototype projects under development. We have also used UA to create an independent portal for clinically oriented data collection, and a custom enrollment interface collections independent of the NIEUW project. As proof of portability, we have hosted the NIEUW related sites on the Heroku[16] cloud platform while the others are hosted on Amazon AWS[17]. We have also dockerized[18] UA to further simplify porting and created a single user mode that allows a user to deploy all portal functions to a single computer, for example a laptop that can be carried into the field. We have tested these by deploying a protected subnet, wholly contained within LDC, for the annotation of sensitive clinical data and deployed a UA portal on a laptop for purposes of demonstration.

In addition to the increased portability, UA has incorporated other features, including: advanced web audio features such as waveforms and recording; integration with cloud storage services such as Amazon S3; a project builder interface that simplifies tool creation for tasks with simpler input, workflow and output requirements. Because the webann and UA code bases are both developed at LDC and remain in use, improvements in one are often shared by the other in a sort of software collaboration. UA is already available to the LREC community in the form of the LanguageARC portal that recruits citizen linguists to work on a variety of data collection and annotation projects. Source code will be released at the project's end.

7. Innovating Language Resource Development

To address the ongoing shortfall of LRs relative to growing need LDC began by surveying alternative approaches to linguistic data collection that tap into renewable sources of the time and effort required. Social media platforms, pro bono efforts such as Librivox which recruits volunteers to read out-of-copyright texts aloud to create audiobooks, and especially citizen science portals such as Zooniverse and Crowd Curio supplement current approaches that use paid experts and paid crowd workers. They demonstrate the nearly boundless power of organizing willing contributors and offering as incentives: challenge, entertainment, competition, opportunities to contribute to technological advancement and by extension the betterment of one's own community and the broader society. For example, Zooniverse has recruited nearly two million citizen scientists who have contributed more than 250 million judgements to researchers in astronomy, biology and other fields.

LanguageARC, built upon LDC's UA framework (see §6), hosts LR development projects and recruits citizen linguists to help them succeed. Each project involves one or more tasks in which a contributor performs a simple activity, once or iterated consistently over multiple items. The activity could involve examining an audio or video clip, image or text and then providing judgements in the form of spoken or written responses, text selections or choices from a controlled vocabulary that are implemented as button clicks. One LanguageARC project documents the variation in British English across the Southeast of England

[14] https://namethatlanguage.org
[15] https://languagearc.com/
[16] https://www.heroku.com/
[17] https://aws.amazon.com/
[18] https://github.com/docker

by asking contributors to judge the region of origin, ethnicity and social class of speakers based on their reading of identical text and by recording themselves saying what they think of the annotation categories used in the prior exercises. Another project investigates linguistic markers of Autism Spectrum Disorders, asking citizen linguists to provide data for psychometric norming by completing language tasks often used in diagnosis: picture description, picture sequence narration, sustained phonation and social attribution of animations. Yet another project elicits dialect terms via picture and silent video description. Still others elicit translations that reveal the grammatical features of the target language. By design, items in LanguageARC tasks take about one minute to complete and tasks can be built in about an hour given the source data and instructional material.

To sustain the citizen linguist community each LanguageARC project presents a title, call to action, project image and pitch. To help contributors connect with research, each project may have partner badges, bios of the research team and discussion forums. Tasks may also have tutorials, reference guides and their own discussion forum.

All resources LDC produces as a direct result of NSF NIEUW will be distributed at the lowest possible costs and with the fewest possible constraints.

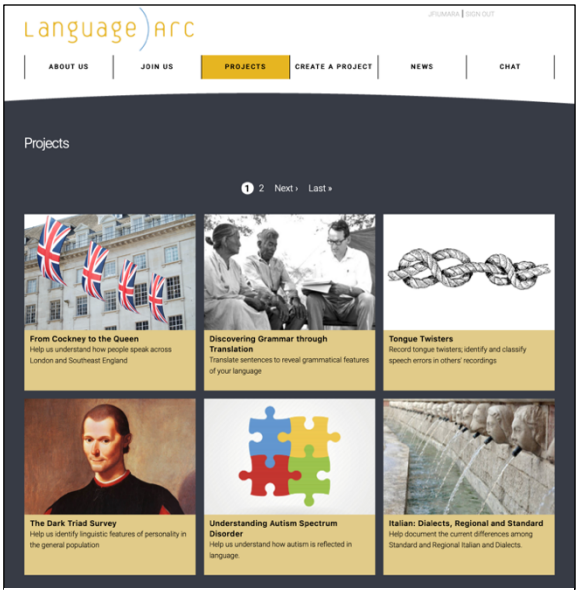

Figure 5: LanguageARC project's menu

8. Conclusion

We have sketched the work the Linguistic Data Consortium has undertaken, since the last report, to support the LREC community including the publication of 96 new corpora, participation in multiple common tasks programs that are already producing resources to the community and will continue to do so over the next several years (the resources, of course will remain accessible) and innovation in methods for data collection and annotation including those that supplement current approaches by offering novel incentives to game players and citizen linguists.

9. Acknowledgements

UA, NameThatLanguage and LanguageARC are made possible by a Research Infrastructure grant (CRI CI-NEW 1730377) from the US National Science Foundation.

10. Bibliographical References

Barker, J., S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in Proc. Interspeech, 2018, pp. 1561–1565.

Bin Li, Xiaopeng Bai, Siqi Yin, Jie Xu (2015) Chinese CogBank: Where to See the Cognitive Features of Chinese Words. Proc. 3rd Workshop on Metaphor in NLP, pp 77–86, Denver, Colorado, June 5.

Christianson, C., Duncan, J. & Onyshkevych, B. (2018) Overview of the DARPA LORELEI Program. Machine Translation 32, pp 3–9.

Cieri, Christopher, Linda Corson, David Graff, Kevin Walker (2007) Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora in Proc. Interspeech 2007, pp 950-953.

Fiscus, J. G., J. Ajot, M. Michel, and J. S. Garofolo (2006) The Rich Transcription 2006 Spring Meeting Recognition Evaluation, In Proc. International Workshop on Machine Learning for Multimodal Interaction. Springer, 2006, pp. 309–322.

Stephan Oepen, Omri Abend, Jan Hajič, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdeňka Urešová, (2019) MRP 2019: Cross-Framework Meaning Representation Parsing. In Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 CoNLL, pages 1–27 Hong Kong, November 3, Association for Computational Linguistics.

Ryant, Neville, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, Mark Liberman (2019) The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In Proc. Interspeech 2019, pp 978-982.

Sadjadi, S.O., Greenberg, C., Singer, E., Reynolds, D., Mason, L., Hernandez-Cordero, J. (2019) The 2018 NIST Speaker Recognition Evaluation. In Proc. Interspeech 2019 pp 1483-1487.

Stephanie Strassel, Dan Adams, Henry Goldberg, Jonathan Herr, Ron Keesing, Daniel Oblinger, Heather Simpson, Robert Schrag, and Jonathan Wright (2010) The DARPA Machine Reading Program-Encouraging Linguistic and Reasoning Research with a Series of Reading Tasks. In Proc. of LREC 2010.

Wright, Jonathan, Kira Griffitt, Joe Ellis, Stephanie Strassel, and Brendan Callahan. 2012. Annotation trees: LDC's customizable, extensible, scalable, annotation infrastructure. In Proc. LREC 2012, pp 479–85.