Reflections on 30 Years of Language Resource Development and Sharing

Christopher Cieri, Mark Liberman, Sunghye Cho, Stephanie Strassel,
James Fiumara, Jonathan Wright
Linguistic Data Consortium
3600 Market Street, Philadelphia, PA. USA 19104
{ccieri, myl, csunghye, strassel, jfiumara, jdwright}@ldc.upenn.edu

Abstract

The Linguistic Data Consortium (LDC) was founded in 1992 to solve the problem that limitations in access to shareable data were impeding progress in Human Language Technology research and development. At the time, the US Defense Advanced Research Projects Agency had adopted the common task research management paradigm to impose additional rigor on their programs by providing shared objectives, data and evaluation methods. Early successes underscored the promise of this paradigm but also the need for a standing infrastructure to host and distribute the shared data. During LDC's initial five year grant, it became clear that the demand for linguistic data could not easily be met by the existing providers and that a dedicated data center could add capacity first for data collection and shortly thereafter for annotation. The expanding purview required expansions of LDC's technical infrastructure including systems support and software development. An open question for the center would be its role in research beyond data development, a question that has since been addressed. Over its 30 years history, LDC has performed multiple roles ranging from neutral, independent data provider to multisite programs, to creator of exploratory data in tight collaboration with system developers, to research group focused on data intensive investigations.

Keywords: language resources, linguistic data, annotation, data centers, data intensive research

1. Introduction

Over its history, the roles of the Linguistic Data Consortium (LDC) [1] have grown in response to, or anticipation of community demands, from publisher, archive and sometime funder of linguistic data created by others, to data collector and annotator, to developer of data specifications and systems for collection and annotation, to research group focused on data intensive investigations. The European Language Resource Association (ELRA)[2] reported a similar expansion in its 20 year retrospective (Choukri et al., 2016). This paper continues a biennial tradition of providing the LREC community with an update on LDC progress but also provides a retrospective view of LDC history and a preview into future plans.

2. Language Data Center Pre-History

Human Language Technology (HLT) research in the mid-1980s in the United States was just beginning to awaken from its long hibernation during the AI winter.

The now infamous Automatic Language Processing Advisory Committee (ALPAC) report (Pierce et al., 1966) had claimed that in the US the demand for translation had been overestimated while the capacity of available human translators in the market was sufficient to meet demand but had been hampered by obstructions in the supply chain. It implicitly rejected the notion that relevant scientific publications should be broadly available to researchers regardless of the languages of those publications' authors and readers. The report determined that readers needed more time to process and understood less from reading machine translation output, even after post-editing which inevitably increased cost over that of direct human translation. The report advised that while the government should support (computational) linguistics as a science, such investments would not yield improvements in machine translation. Instead, it recommended investment into measuring the quality of – and developing computer systems to support – human translations. As noted elsewhere, the ALPAC report: *"virtually ended MT research in the US for more than a decade"* in favor of systems to assist human translators (Hutchins, 2001).

Just 3 years later, the ALPAC report's lead author (Pierce, 1969) wrote, in the Journal of the Acoustical Society of America's Letters to the Editor, a similarly damaging criticism of speech recognition research. In that letter, one again sees the laudable desire to focus effort on language science and measurable outcomes coupled with a failure to imagine paths toward sufficient improvement of the technology or its possible applications which he called: *"as specious as insisting that an automobile should respond to gee"* before noting *"we do very well with keyboards, cards, tapes, and cathode-ray tubes"*. The letter also had a deleterious effect on speech technologies' funding (Church, 2017) though some work, for example the Speech Understanding Research (SUR) program, continued through the mid 1970s before the Defense Advanced Research Projects Agency (DARPA) funding of such research took a 10 year hiatus (Liberman and Wayne, 2020).

The reawakening of HLT in the U.S. that has led to the past ~35 years of intensive work, with strong government support, began around 1986 with the new DARPA Speech Recognition program, which focused on both speaker dependent and independent read speech, within restricted domains, with vocabularies ranging from 1k to 64k words.

DARPA HLT programs from this point forward employed the *common task* research management paradigm within which multiple teams work in parallel, cooperating and competing, to achieve well defined, quantifiable goals

[1] https://www.ldc.upenn.edu
[2] http://www.elra.info/en/

using shared data with regular (also frequent) evaluation of system performance administered by a neutral party using objective, pre-determined criteria. A normal cycle of such programs culminates in a workshop where program members discuss objectives, challenges, data, research approaches, results and even the evaluation criteria, making course corrections as needed and then continuing the virtuous cycle until program goals are reached or funding is spent (Liberman and Wayne, 2020). Of particular value is that performers are required to describe their innovations and the impact of innovations on system performance so that all may adopt winning approaches and avoid pitfalls. Such programs create a culture of knowledge and resource sharing that attracts new research to problems of interest even in the absence of funding (Church, 2017).

By the early 1990s, as a result of DARPA successes, HLT research was described optimistically as having *"useful present-day systems and realistic expectations of progress"* though the extent to which Language Resources (LR) or their absence or limitations could affect progress remained a foremost concern: *"Not even the largest companies can easily afford enough of [the needed] data to satisfy their research and development needs"* and *"Researchers at smaller companies and in universities risk being frozen out of the process almost entirely"* (Liberman and Godfrey, 1993). Researchers in Europe also recognized "*growing worldwide awareness of the need for […] publicly available common corpora*" (Gibbon, Moore, and Winski, 1998).

The solution proposed to this quandary was to enable resource sharing on a scale adequate to meet current needs through a mechanism that could also anticipate future needs. At that time, such LRs that existed were closely held by their creators or else distributed unevenly under inconsistent terms and timelines. Allowing that situation to continue would have reinforced the schism between haves and have-nots and impeded the progress that the *common task* methodology was meant to encourage, progress that we have luckily observed over the past 30 years.

DARPA recognized the need for an organization to focus on acquiring, curating and distributing LRs to support research education and technology development. Although research groups were and are capable of performing these functions, DARPA and the community saw value in centralizing the distribution function as well as the technologies (e.g. CD-ROM production at the time) and skills required in order to improve quality and reduce cost through the effects of greater scale. Through an open call for proposals, the University of Pennsylvania was selected as host organization and, with DARPA support, *"the Linguistic Data Consortium was founded in 1992 to provide a mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies"* (Liberman and Godfrey, 1993). The US National Science Foundation (NSF) and National Institute of Standards and Technology (NIST) would add support in LDC's first years and many other government branches have offered support subsequently.

3. Data Distribution

LDC's early efforts were focused on: "*distributing previously created datasets, and funding or co-funding the development of new ones*" (Liberman and Godfrey, 1993). The latter was enabled by the seed grant from DARPA which included the requirement that LDC become self-supporting through membership and data licensing fees. This insistence on sustainability has allowed LDC to support multiple research communities over the intervening 30 years. Between 1992 and 1995, LDC focused exclusively on corpus distribution. An advisory board including members from the non-profit, government and commercial sectors defined the LDC business model that is still effect today with small adaptations. LDC is a Consortium, a kind of mutual aid society where members provide support in the form of membership fees and data contributions and receive in return access to many, many more datasets than any one member could hope to create. Although LDC is also able to individually license many of the corpora in its Catalog, membership continues to be the best option. In short, organizations contribute an annual fee and receive free ongoing access to the corpora released during each year of their membership. Organizations that embraced the consortium model reported being the most satisfied, 95% on average across multiple membership surveys (Reed, DiPersio, and Cieri, 2008).

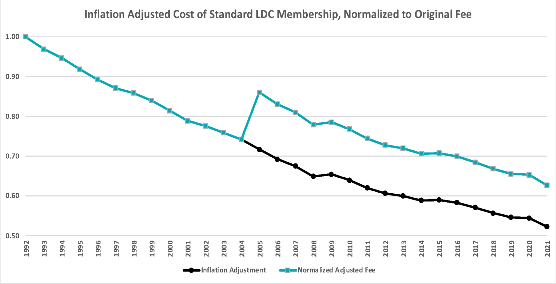


Figure 1: Decreasing cost of LDC Membership due to its stability relative to inflation

Over LDC's 30 year history, standard membership fees were raised once; at the same time LDC provided a new membership type to support larger organizations who needed all LDC corpora as released. Over the same period, inflation has reduced the 'buying power' of a dollar in 2021 to just 52% of what it was in 1992. The stability of the membership fees relative to inflation alone yields increasing value over time as shown in Figure 1 where the inflation adjusted fee of a standard membership, even including the 2005 increase, is now just 63% of what it was in 1992.

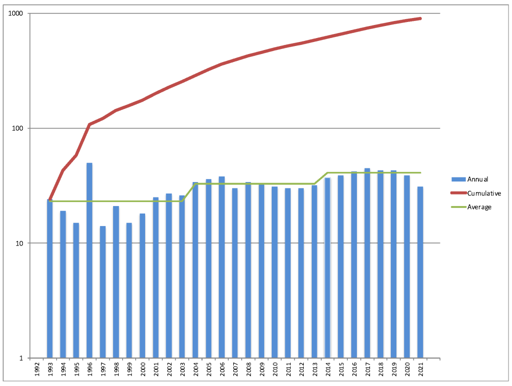


Figure 2: Increasing value of LDC Membership by the number of corpora published per year, 10 year average and cumulative

While inflation continues to reduce the relative cost of LDC memberships, their value in terms corpora released per year, corpus size and diversity continues to rise. In LDC's first eight years the number of corpora released annually was highly variable (mean=18, std=11.8). We began a concerted effort between 2001 and 2003 to stabilize practice to ~2 corpora per month (annual mean=26, std=1). Beginning in 2004, the target was increased to ~30 per year (mean=36, std=5). Although a metric such as corpora per unit cost does not comprehend differences in corpus source data, volume and annotation complexity, etc., the nearly halving of the buying power of the dollar coupled with the doubling of the average number of available corpora per year is compelling evidence that a dedicated language data center can exploit economies of scale and increasing efficiencies in technology and process to offer continually growing benefit to its members.

Today's LDC Catalog includes more than 900 corpora in 107 linguistic varieties, including recent additions in Dari, Georgian, Icelandic, Kazakh, Kurdish, Nahuatl, Persian, Pushto, Russian, Turkish Ukrainian, Uzbek and Zulu.

These corpora were developed and/or used within 91 research programs including the following:

- large multisite programs sponsored by DARPA, IARPA and other agencies: AIDA, AQUAINT, BEST, BOLT, Communicator, DEFT, EARS, GALE, HARD, HAVIC, Hub4, Hub5-LVCSR, KAIROS, LCTL, LORELEI, Machine Reading, MADCAT, MED, RATS, ROAR, SPINE, TDT, TIDES, Tipster, Transtac
- NIST evaluation campaigns: LRE, ACE, MT, OpenHaRT, OpenSAD, OpenSAT, RT, SRE, TAC/KBP, TREC, TRECVid
- community organized evaluations: CoNLL, SemEval, SIGHAN

The LDC Catalog[3] lists 75 target applications of LDC data the most common of which are:

- handwriting recognition
- entity, event, relation extraction & coreference
- information retrieval
- knowledge base population
- language identification
- language modeling
- machine translation
- parsing, POS tagging & other NLP
- pronunciation modeling
- question-answering
- semantic role labelling
- sentiment detection
- speaker diarization, identification
- speech activity detection
- speech recognition
- summarization

4. Data Collection and Annotation

By 1995 it had become clear that language data collection at the scale demanded by contemporaneous DARPA programs required teams dedicated to corpus development and supplied with specialists and specialized technologies. LDC collection effort began in 1995, focused on conversational telephone speech to meet imminent program deadlines. Over the intervening years, operations expanded to include regular collection of newswire and broadcast news, transcription and the development of pronouncing lexicons. All of this new data collected by LDC is, of course, added to the Catalog.

In response to program needs over the past three decades, LDC has locally implemented or else innovated methods for collecting text from:

- news sources, journals, financial and biomedical documents
- internet sources including newsgroups, blogs, microblogs, comment threads and discussion forums
- text interactions via email, chat and SMS
- scans or images of documents containing printed or handwritten text or both.

Among audiovisual data types we have collected:

- broadcast news and conversation, podcasts
- conversational telephone speech
- lectures, interviews, meetings, field interviews
- read, prompted & task oriented speech, role play
- speech in noise
- web video and directly contributed amateur video
- animal vocalizations
- digitized analog media including interviews in a variety of tape formats
- two way radio speech characterized by severe channel noise

Some of these efforts required creation of new hardware software solutions including:

- satellite downlink node on the Voice of America network to collect multilingual broadcast news (Cieri and Liberman 2000)
- fully automated platform for collecting broadcast audio and video and processing those through ASR and MT technologies
- subsequent miniaturization of the platform to enable outsourcing of collection to international partners (Walker, Caruso and DiPersio 2010)
- platforms to collect telephone calls also eventually miniaturized, modularized, deployed and managed remotely
- interfaces that allow users to upload their messaging archive and remove any sensitive messages before contributing the remainder.
- digitization station that can accept input from most common legacy analog media players
- platform to broadcast and receive, and optionally degrade, clean audio for which we had transcripts

It has also been necessary to innovate new annotation tools, often highly customized, task-specific local tools (Maeda, Mazzucchi, Cieri and 2011). However, over the past 10 years, LDC has shifted nearly all annotation tasks to a web based tool, *LDC webann* (Wright et al. 2012), that presents customized front ends, that appear to users as distinct tools but use the same underlying database schema and project

[3] https://catalog.ldc.upenn.edu

management umbrella. This innovation allowed LDC's annotation team to be remote if necessary, and when it became necessary in March 2020, no technological adjustment was necessary. The database supporting the application recognizes the evolving tools with new records representing the logical form of the data, and flat, unnormalized, tables more suitable for querying the data can be constructed automatically on an as-needed basis.

Quantifying the amount of annotation webann has supported is challenging as different annotation tools require and package decisions at different granularities but, conservatively estimated, ~1000 annotators have used ~1000 webann tasks to contribute ~10 million annotations.

While developed for common task programs, as LDC needs expanded beyond the typical bounds of such programs, webann evolved into a new application, called Universal Annotator (UA) with a much more flexible deployment design, supplying the basis for the NIEUW portals (see §4.2) and *webtrans*, a comprehensive audio transcription application.

Abstracting from the details of individual programs and their needs, longer term data development efforts have often followed one or more of these trajectories:

- steadily increasing demand for greater volume at lower cost with shorter timelines
- steadily increasing diversity of data sources, domains, genres for a given language
- initial focus on data quantity followed by a reassessment of the role of data quality
- initial focus on familiar languages followed by increased focus on linguistic diversity whether accomplished by increasing the number of languages in response to emergent need or by methodically sampling language families or typological features
- increasing specificity in the linguistic varieties sampled

4.1 Recent Data Collection & Annotation Efforts

A complete accounting of the dozens of data collection and annotation efforts and their complex characteristics would be out of scope for this paper. However, a sampling of some very recent efforts indicates the breadth of current work and the LRs that will be available to the research community as a result.

We Can Talk is a multi-modal, multi-lingual speech corpus for speaker recognition, consisting of telephone speech and speech from video for over 200 multilingual speakers of Cantonese plus Mandarin and/or English.

The *Corpus of Annotated Multilingual Images* for OCR (CAMIO) supports optical character recognition research and evaluation, containing data from 35 languages across 24 unique scripts. The corpus comprises approximately 70,000 images, most of which have been annotated for text localization resulting in over 2.3M bounding boxes around lines of printed text. For 13 of the languages, a subset of ~16,000 images have been transcribed, yielding over 2.4M tokens of text data.

The *Kurmanji and Sorani Speech Transcripts* (KASET) contains ~ 350 hours of audio data from broadcasts and conversational telephone speech recordings in the two Kurdish varieties, plus transcripts for approximately 65 hours of audio.

The *Corpus of Speech in Natural Environments* (COSINE) corpus contains ~500 hours of audio from multiple genres in five languages: Indonesian, Korean, Mandarin, Modern Standard Arabic (MSA), and Russian, plus ~300 hours of orthographic transcription and ~75 hours of translation of the audio recordings.

The *AIDA (Active Interpretation of Disparate Alternatives (AIDA)* corpora contain thousands of multimedia documents (text, image and video) in Russian, Ukrainian, Spanish and English covering several current event scenarios like Russia-Ukraine relations and the COVID pandemic, with a particular focus on documents reflecting disinformation and conflicting claims about the same situation. AIDA annotations include entities, relations and events in a large ontology, with cross-document coreference, association of labeled events with the specific claims they comprise, and an indication of claims that support, refute and relate to one another.

KAIROS (Knowledge-directed Artificial Intelligence Reasoning Over Schemas) corpora include the ~15M-document Schema Learning Corpus, designed to enable induction of high-level representations (i.e. schemas) for complex events like purchases, bombings or travel, across a broad set of domains. Each complex event type includes a proto-schema describing the typical steps for that event, plus labeled examples of each step present in the Spanish/English multimedia corpus. KAIROS evaluation corpora include documents related to real-world IED attacks, disease outbreaks and other incidents, with annotation of the temporally-ordered steps associated with each incident along with the entities, relations and events comprising those steps.

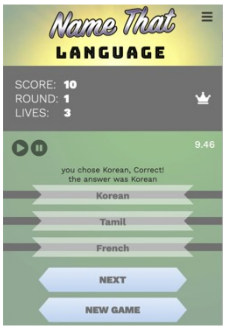
In addition to these new efforts, LDC has undertaken additional work on corpora presented at previous LRECs, for example, the transcription and translation of the *CallMyNet2* (CMN2) corpus. The primary objective of CMN2 transcription was to produce orthographic transcripts for 400 hours of Tunisian Arabic conversational telephone speech recordings selected from CMN2, focusing on data utilized in the NIST SRE18 evaluation. Transcripts represent a content-accurate, limited markup orthographic transcript using a transcription system based on Modern Standard Arabic orthography but reflecting Tunisian Arabic pronunciation. LDC then translated ~200 hours of the transcribed conversations; data selected for translation reflected data from the full set of claque speakers in the original CMN2 corpus.

4.2 Novel Incentives

The recognition that years of concentrated effort by many data centers, national language projects and international consortiums have not met LR demand has motivated LDC to investigate *novel incentives*. As part of a three year program (subsequently extended due to the negative impact of the pandemic on outreach), LDC has built three portals and collected a variety of raw data and annotations and supported collaborators doing the same. *NIEUW* (Novel

Incentives, Workforces and Workflows) is supported by NSF's Computer & Information Science & Engineering Research Infrastructure program and uses novel incentives to supplement more familiar efforts using monetary compensation.

While the HLT community continues to suffer LR shortages, online communities have shown that people are willing to spend time providing language data given sufficient motivation and appropriate tools. Initiatives such as LibriVox elicit massive donations of time and spoken data by appealing to contributors' interests and altruism. LibriVox readers record themselves reading works of literature (out of copyright, at least in the US) aloud to create public domain audiobooks. Contributors respond to a range of motivations: a philanthropical drive to improve access to literature in general – or the works of a particular author or a single work – among children, the sight impaired or illiterate; a desire to join a community of like-minded people; the ambition to improve upon prior versions; the need to develop their own skills as voice actors and/or to gain recognition and perhaps even paid work from other sources (Some LibriVox readers also read for Audible.com).



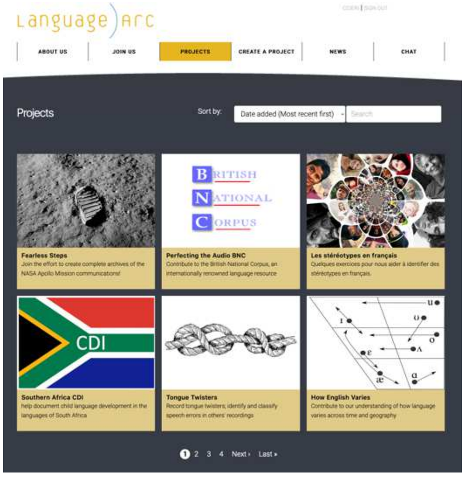
SCORE: 10
ROUND: 1
LIVES: 3

you chose Korean, Correct!
the answer was Korean

Figure 4: Eliciting language judgments via novel incentives: entertainment and competition

NIEUW began by developing a game, NameThatLanguage[4] which elicits judgements of the languages spoken in short audio clips. Even though many of the players are non-experts, their judgements can be aggregated to predict the correct answer with very high confidence. The input to the game was a set of >=80 clips for each of 13 languages plus 5400 clips suspected to be in one of 9 languages. To date, the game has collected the results of ~720,000 HITs presented to ~46,000 unique player IDs who provided useable results for 86%. Aggregated results converge on the correct answer for known clips after different numbers of responses depending upon the language. However, the player pool has achieved confidence sources of >=.98 for nearly all clips. When the player pool does not converge on the expected answer this nearly always (96%) means the clip is not in the suspected language (Cieri, Fiumara, and Wright, 2021).

The second phase of the NIEUW program developed a portal, LanguageARC[5] where Citizen Linguists can contribute to a variety of data intensive projects (Fiumara, et al., 2020). In one recent project, ~59 native speakers of the Xi'an Guanzhong variety of Mandarin recorded themselves naming objects appearing in 622 images from the MultiPic corpus (Duñabeitia et al., 2018) selected for their familiarity to people living in China. A subset of these contributors audited each of 34,729 recordings for audio quality and use of the target variety. More recently LDC and researchers at the University of Texas, Dallas have begun to elicit transcriptions of the extremely challenging audio in the Fearless Steps (Joglekar et al., 2021) corpora containing communications of NASA Apollo space missions. Although contributors may be entertained or seek competition in Citizen Linguistics projects, principal incentives are most likely opportunities to learn, contribute to social good or reinforce local pride, for example by documenting an under-represented variety.



Sort by: Date added (Most recent first)

Fearless Steps
Join the effort to create complete archives of the NASA Apollo Mission communications!

Perfecting the Audio BNC
Contribute to the British National Corpus, an internationally renowned language resource

Les stéréotypes en français
Quelques exercices pour nous aider à identifier des stéréotypes en français.

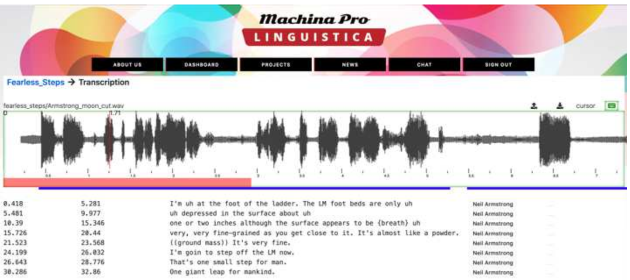
Southern Africa CDI
help document child language development in the languages of South Africa

Tongue Twisters
Record tongue twisters, identify and classify speech errors in others' recordings

How English Varies
Contribute to our understanding of how language varies across time and geography

Figure 3: LanguageARC eliciting data and judgements through novel incentives: learning, altruism, local pride

Finally, NIEUW has developed *Machina Pro Linguistica*, (hereafter MachPro)[6] for language professionals, linguists and others who work with language regularly, and may be motivated to contribute as a way to develop their professional skills, supplement their learning or gain access to resources in exchange.



fearless_steps/Armstrong_moon_cut.wav

```
0.418    5.281    I'm uh at the foot of the ladder. The LM foot beds are only uh       Neil Armstrong
5.481    9.977    uh depressed in the surface about uh                                 Neil Armstrong
10.39    15.346   one or two inches although the surface appears to be {breath} uh     Neil Armstrong
15.726   20.44    very, very fine-grained as you get close to it. It's almost like a powder.  Neil Armstrong
21.523   23.568   ((ground mass)) It's very fine.                                      Neil Armstrong
24.199   26.032   I'm goin to step off the LM now.                                     Neil Armstrong
26.643   28.776   That's one small step for man.                                       Neil Armstrong
30.286   32.86    One giant leap for mankind.                                          Neil Armstrong
```

Figure 5: MachPro eliciting data through novel incentives: skills development, access to resources, supplemental learning.

In one project, contributors transcribe in exchange for access to the Penn Sociolinguistic Archive, containing more than 5800 recordings collected over a span of more than 50 years by Professor William Labov and his students. MachPro contains all of the tool building capabilities of LanguageARC as well as a full implementation of LDC webtrans, the tool used within LDC for recent transcription projects. In addition, MachPro project designers have greater flexibility in creating documents to explain

[4] https://namethatlanguage.org
[5] https://languagearc.com
[6] https://machprolx.org/

annotation methods and to link among documents and between documents and tools. While Citizen Linguistics projects in LanguageARC are open by default, users of MachPro can easily create projects that are private to the students in a class or members of a research team.

The presence, within the same organization, of researchers working on novel incentives and the language of clinical interactions, described below, led to the creation of SpeechBiomarkers.org, where volunteers can do brief exercises such as picture descriptions to help establish baseline population performance for research involving clinical populations.

This use of novel incentives offers an opportunity to expand beyond the limits of funded programs and begin to document and provide data for technology development in languages that have heretofore remained under-served.

5. Data Intensive Research

An open question that every data center must address is the appropriate role for data creators in the research ecosystem. Here we begin by noting that, notwithstanding common parlance, the boundary between data collection and research is illusory. Data collection is not some necessary evil that must be endured before research can begin but is rather an early and integral phase of the research endeavor. Decisions made during data collection and annotation can impact, e.g. system performance by failing to create an appropriate sample, by inserting bias, or by introducing annotation errors that can be mistaken for system errors in a technology evaluation. That said, there remain multiple roles that data centers can fill depending on the nature of their research communities and the abilities of center staff.

Within the context of the DARPA common task programs described earlier, LDC was often required to play the role of neutral provider of both training and test data. In that role, LDC project managers needed to understand the multiple needs of the sponsors, system developers and evaluation teams as well as the limitations of any existing data and human annotators. In such a role, for example, pre-annotation using any system – but especially any system developed by one of the performers – had to be done carefully, if it were done at all, to avoid biasing the data in favor of that performer. Furthermore, the data creation effort had to be insulated from any related research such as LDC's own work on translingual topic tracking during the Topic Detection and Tracking program (Schultz and Liberman, 1999). Finally, that role brings considerations other than maximizing the performance of systems built upon the training data and evaluated with the test data. Sponsors often want the data to be increasingly realistic over time which can lead to temporary degradations of performance. At the same time, evaluation teams want test data of sufficient size and diversity to yield statistically significant results; however in a program with a fixed data budget, funds devoted toward test data reduce funds available for training and development data.

In other programs, especially in their early stages or when specifications were not yet mature, LDC was required to provide exploratory data and annotations, making many small distributions and reacting to feedback. In such cases, a tighter coupling between data creation and system development was acceptable.

A third potential role for data centers includes direct involvement in all aspects of research which may include technology development and evaluation (Choukri et al 2016).

As an example of the latter case, LDC has organized multiple evaluations of diarization technology (DIHARD), selecting and annotating data to present new challenges to performers, contributing to the discussion of evaluation metrics and, when necessary, running the scorers and reporting results to performers. The program's contributions were summarized neatly in a recent report-out: "*The evaluation results indicate that speaker diarization has improved markedly since DIHARD I, particularly for two-party interactions, but that for many domains (e.g., web video) the problem remains far from solved.*" (Ryant et al., 2021). Running technology evaluations on this scale requires skills in short supply in many research groups though abundant at NIST. LDC has been grateful for NIST interest in the latest DIHARD campaigns.

Examples of LDC direct involvement in research and development abound in the clinical domain where we were able to find partners who were open to collaboration seeking linguistic biomarkers of the conditions they study and willing to adapt their data collection processes as needed. Other data centers and data intensive projects have similarly begun to focus on "atypical speech" (van der Heuvel et al. 2020)

A series of collaboration projects with the Center for Autism Research (CAR) at the Children's Hospital of Philadelphia has explored the possibility of speech as robust, quick, and low-cost biomarkers of autism. Early collaboration with the Center for Autism Research (CAR) and the Children's Hospital of Philadelphia showed that language and speech features drawn from clinical interviews with expert clinicians provide strong, sensitive biomarkers for distinguishing autistic children from neurotypical children (Parish-Morris et al., 2016a, 2016b). Another paper with CAR (Parish-Morris et al., 2017) shows an interesting finding that autistic girls sounded more typical in their use of pause fillers ("um") during natural conversations than did autistic boys and the increased "um" ratio might be autistic girls' strategy for "linguistic camouflage". Since our previous studies reported that language and speech features from clinical interviews were valuable in understanding and prescreening autism, we further explored if this still held true when conversations were not led by expert clinicians. In a follow-up study (Cho et al., 2019), which examined this possibility, we were able to show that language and speech features extracted from very brief, natural conversations with naïve interlocutors (about 5 minutes long) could be indeed used for automatically identifying autistic children, suggesting that prescreening using speech may be possible in a non-clinical, local community setting, including historically marginalized or low-resource communities.

Another ongoing, active collaboration with the Frontotemporal Degeneration Center at the University of

Pennsylvania Hospital investigates the intricate links between neurodegeneration, speech performance, and underlying pathology. In multiple reports, we were able to show that language and speech features extracted from one-minute picture description data using automated acoustic and lexical pipelines were useful in better understanding various neurodegenerative conditions, including frontotemporal dementia (Nevler et al., 2017, 2019; Cho et al., 2021a), amyotrophic lateral sclerosis spectrum disorders (Nelver et al., 2020), progressive supranuclear palsy and corticobasal syndromes (Parjane et al., 2021), and Alzheimer's dementia (Cho et al., 2021b). Detailed explanation of the acoustic pipeline (Nevler et al., 2017), the lexical pipeline (Cho et al., 2021a) and applications of these pipelines to the study of healthy aging populations (Cho, Nevler et al., 2021) are published elsewhere. We also showed that the language and speech features from our automated pipelines were useful in automatically identifying patient groups with different neurodegenerative phenotypes or pathology (Cho et al., 2020, Cho et al., in press). Since a short and simple speech sample, such as the one-minute picture description task, had success in understanding various types of neurodegenerative disease, our recent work extends beyond picture descriptions, exploring whether other types of frequently administered neuropsychological tests, such as fluency tasks or story recall tasks, would help in characterizing different phenotypes and pathology in patients with neurodegeneration. We established an automatic processing pipeline for letter-guided fluency tasks (Cho et al., 2021d) and the pipeline for story recall data is under development. We are currently collecting patients' data using in-person and remote data collection methods, and we plan to apply the pipelines to patients' data soon.

A collaboration with researchers from the Feinstein Institutes at Northwell Health examine language, psychosis, and social processing using cutting-edge technologies. In previous work, we showed that language features from clinical interviews had greater ability to discriminate between patients with schizophrenia spectrum disorder and healthy speakers (Tang, Kriz, Cho et al., 2021) than traditional clinical ratings for psychosis. In recent works, we are studying the relation between various psychotic symptoms, social/emotion processing and language features (e.g., Krell, Tang, et al., 2022).

It appears that the application of data collection and annotation methodologies, feature extraction pipelines and analytic processes developed in many of the programs described above have great potential to advance research, not only in clinical disciplines but also in a wider range of disciplines that rely upon linguistic data.

6. Conclusion

We have reviewed LDC's roles over the past 30 years after describing the conditions that lead to an HLT winter followed by a reawakening and an insatiable hunger for LRs. We have described data distribution, collection and annotation and other research with somewhat less individual detail than in previous reports, opting instead to provide a longer view dotted with occasional specifics. Where possible, we have tried to raise issues that strike us as relevant to all data centers, the communities they serve and the future of our allied research.

7. Acknowledgements

For a paper with this scope it is, unfortunately, impossible to name all of the individuals and organizations that have contributed to LDC's progress. Nonetheless, we would like to express our gratitude to the agencies that fund LR development, particularly, DARPA, NIST and the National Science Foundation (including CRI grant 1730377 which supports the NIEUW project described above), who were the first to support LDC. We would also like to thank the hundreds of individual researchers in the many disciplines that rely upon linguistic data who have collaborated with us and enriched our joint work. Most especially we would like to thank the thousands of LDC members, data users and data contributors who have invested in the Consortium over its history.

8. Bibliographical References

Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R., and Parish-Morris J. (2019). Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. *Proc. Interspeech*, 2513-2517.

Cho, S., Nevler, N., Ash, S., Shellikeri, S., Irwin, D., Massimo, L., Rascovsky, K., Olm, C., Grossman, M., and M. Liberman. (2021a). Automated analysis of lexical features in frontotemporal degeneration. *Cortex* 137: 215-231.

Cho, S., Cousins, K., Shellikeri, S., Ash, S., Irwin, D., Liberman, M., Grossman, M., and Nevler, N. (in press). Lexical and acoustic speech features relating to Alzheimer's disease pathology. https://www.medrxiv.org/content/10.1101/2021.09.27.21264148v2 (preprint on medRxiv)

Cho, S., Nevler, N., Shellikeri, S., Parjane, N., Irwin, D., Ryant, N., Ash, S., Cieri, C., Liberman, M., and Grossman, M. (2021). Lexical and acoustic characteristics of young and older healthy adults. *Journal of Speech, Language and Hearing Research* 64(2):302-314.

Cho, S., Shellikeri S., Ash, A., Grossman ,M., Nevler, N., and Liberman, M. (2020). Automatic classification of primary progressive aphasia patients using lexical and acoustic features. *Proc. 12th Language Resources and Evaluation Conference 2020 workshop on Resources and Processing of linguistic, para-linguistics, and extra-linguistic data from people with various forms of cognitive, psychiatric, and/or developmental impairments* (RaPID-3).

Cho, S., Shellikeri, S., Ash, S., Liberman, M., Grossman, M., and Nevler, N. (2021c). Automatic classification of AD versus FTLD pathology using speech analysis in a biologically confirmed cohort. Alzheimer's & Dementia: the *Journal of the Alzheimer's Association* 17(S5).

Cho, S., Nevler, N., Parjane, N., Cieri, C., Liberman, M., Grossman, M., and Cousins, K. (2021d). Automated analysis of digitized letter fluency data. *Frontiers in Psychology* 12, 654214.

Choukri, K. Mapelli, V., Mazo, H., and Popescu, V. (2016). ELRA Activities and Services. *Proc. 10th International Conference on Language Resources and Evaluation: 463-468.*

Church, K. W. (2017). Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering*, 24(1): 155–160.

Cieri, C. and Liberman M. (2000) Issues in Corpus Creation and Distribution: The Evolution of the Linguistic Data Consortium, Proc. 2nd International Conference on Language Resources & Evaluation.

Cieri, C., Fiumara, J., and Wright, J. (2021). Using Games to Augment Corpora for Language Recognition and Confusability, *Proc. 22nd Annual Conference of the International Speech Communication Association (Interspeech*), August 30-September 3.

Duñabeitia, J.A, Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., and Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*. 71(4):808-816.

Fiumara, J., Cieri, C., Wright, J., and Liberman, M. (2020). LanguageARC: Developing Language Resources Through Citizen Linguistics in *Proc. 12th Edition of the Language Resources and Evaluation Conference (LREC)*. CLLRD Workshop: Citizen Linguistics in Language Resource Development. Marseille, May 11-16.

Gibbon, D., Moore, R., Winski, R. eds. (1998). Spoken Language Reference Materials. Vol. 4. Walter de Gruyter.

Hutchins, W. J. (2001). Machine translation over fifty years, *Histoire, Epistemologie, Langage*, 22(1):7-31.

Joglekar, A, Seyed, O. S., Chandra-Shekar, M., Cieri, C., and Hansen, J.H.L. (2021). Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data Across NASA Apollo Audio in *Proc. 22nd Annual Conference of the International Speech Communication Association (Interspeech)*, August 30-September 3.

Krell, R., Tang, W., Hänsel, K., Sobolev, M., Cho, S., Berretta, S., and Tang, S. (2022). Lexical and acoustic correlates of clinical speech disturbance in schizophrenia. In Sharban-Nejad, A., Michalowski M., and Bianco, S. (eds.), AI for Disease Surveillance and Pandemic Intelligence. W3PHAI 2021. *Studies in Computational Intelligence,* vol. 1013. Springer, Cham.

Liberman, M. and Godfrey, J. (1993). The Linguistic Data Consortium. In Chen, Keh-Jiann, Chu-Ren Huang, *Proc. ROCLing Computational Linguistics Conference VI*, Nantou, Taiwan, September. Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Liberman, M. and Wayne, C. (2020). Human Language Technology, AI Magazine , 41(2):22-35.

Maeda, K., Mazzucchi, A. and Cieri, C. (2011) Technical Infrastructure Supporting Large-scale Linguistic Resource Creation in Olive, J. Christianson, C. and McCary, J. eds, Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, New York, Springer.

Nevler, N., Ash, S., Jester, C., Irwin, D., Liberman, M., and Grossman, M. (2017). Automatic measurement of prosody in behavioral variant FTD. *Neurology* 89:1-7.

Nevler, N., Ash, S., Irwin, D., Liberman, M., and Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology* 6:4-14.

Nevler, N., Ash, S., McMillan, C., Elman, L., McCluskey, L., Irwin, D., Cho, S., Liberman, M., and Grossman, M. (2020). Auomated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. *Neurology* 95(12): e1629-e1639.

Parish-Morris, J., Liberman, M., Ryant, N., Cieri, C., Bateman, L., Ferguson, E., and Schultz, R. (2016a). Exploring autism spectrum disorders using HLT. *Proc. 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 74-84.

Parish-Morris, J., Cieri, C., Liberman, M., Bateman, L., Ferguson, E., and Schultz, R. (2016b). Building language resources for exploring autism spectrum disorders. *Proc. 10th International Conference on Language Resources and Evaluation*, 2100-2107.

Parish-Morris, J., Liberman, M., Cieri, C., Herrington, D., Yerys, B., Bateman, L., Donaher, J., Ferguson, E., Pandey J., and Schultz, R. (2017). Linguistic camouflage in girls with autism spectrum disorder. *Molecular Autism* 8, 48.

Parjane, N., Cho, S., Ash, S., Cousins, K., Shellikeri, S., Liberman, M., Shaw, L., Irwin, D., and Grossman, M. (2021). Digital speech analysis in progressive supranuclear palsy and corticobasal syndromes. *Journal of Alzheimer's Disease* 82:33-45.

Pierce, J. R., J. B. Carroll, E. B. Hamp, D. G. Hays, C. F. Hockett, A. G. Oettinger, and A. Perlis. (1966). Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC.

Pierce, J. R. (1969). Whither Speech Recognition? *The Journal of the Acoustical Society of America* 46:1049.

Reed, M., DiPersio, D., and Cieri, C. (2008). The Linguistic Data Consortium Member Survey: Purpose, Execution and Results. *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, May 28-30.

Schultz, J.M. and Liberman, M. (1999). Topic Detection and Tracking using idf-weighted Cosine Coefficient, *Proc. DARPA Broadcast News Workshop*.

Tang, S., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R., Bhati, M., Sedoc, J., and Liberman, M. (2021) Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophrenia* 7, 25.

van den Heuvel, H, Oostdijk, N., Rowland, C., and Trilsbeek, P. (2020), The CLARIN Knowledge Centre for Atypical Communication Expertise, Proc. 12th Language Resources and Evaluation Conference, pp. 3312—3316.

Walker, K., Caruso, C. and DiPersio, D. (2010) Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development, Proc. 7th International Conference on Language Resources and Evaluation.

Wright, J., Griffitt, K., Ellis, J., Strassel, S., Callahan, C. (2012) Annotation Trees: LDC's Customizable, Extensible, Scalable Annotation Infrastructure, Proc. 8th International Conference on Language Resources and Evaluation.

9. Language Resource References

Canavan, A., Zipperlen, G. Graff, D. (2014). CALLFRIEND Farsi Second Edition Speech LDC2014S01. Philadelphia: Linguistic Data Consortium.