The Linguistic Data Consortium: Developing and Distributing Language Resources4All

Denise DiPersio, Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
dipersio@ldc.upenn.edu, ccieri@ldc.upenn.edu

Abstract

The Linguistic Data Consortium (LDC) is an open collective of academic, government and industry organizations whose mission is to support language-related research, education and technology development by creating and sharing resources, such as data, tools and standards. Its online catalog is a rich, curated repository of speech, text, video and lexical data sets. LDC develops and publishes resources in a growing number of underserved languages. This paper examines relevant LDC corpora and the "language pack" data set model as successes in resource creation, along with the Consortium's involvement in efforts that advance access to data for all language communities.

Keywords: language resources, digital repositories, language communities

Résumé

Ilé-iṣẹ́ Àgbájọ Fún Àkójọpọ̀-ọ̀rọ̀-èdè Ajẹmọ́-ìmọ̀-ẹ̀dá-èdè (Linguistic Data Consortium (LDC)) jẹ́ ilé-iṣẹ́ àgbájọ àwọn onímọ̀-ìjìnlẹ̀, àwọn ìjọba àti àwọn ilé-iṣẹ́ ìmọ̀ ẹ̀rọ, pẹ̀lú èròǹgbà láti ṣe àtìlẹ́yìn fún ìdàgbàsókè ìjìnlẹ̀ ìwádìí ajẹmọ́-èdè, èto ẹ̀kọ́, àti ìmọ̀ ẹ̀rọ, nípa pípèsè àti pípín àkójọ-èdè, gẹ́gẹ́ bí àkójọ-ọ̀rọ̀ nínú èdè, àwọn oríṣiríṣi ohun-èèlò, àti àwọn ìlànà ìgbéléwọ̀n. Ààtò àgbéjáde rẹ̀ nínú ẹ̀rọ ayélujára kún fún àkópamọ́ ọ̀rọ̀-ẹnu geere, ọ̀rọ̀-àkọsílẹ̀ oníhun, àti oríṣiríṣi àká-ọ̀rọ̀. Ilé-iṣẹ́ yìí (LDC) ń ṣe iṣẹ́ ìdàgbàsókè pẹ̀lú àwọn àtẹ̀jáde, lórí ìpèsé ohun àmúlò fún ọ̀pọ̀lọ́pọ̀ àwọn èdè kéékèèkéé tí wọ́n ń dàgbàá bọ̀. Ìwé àpilẹ̀kọ yìí ń ṣe àlàyé lórí àwọn àkójọpọ̀ ọ̀rọ̀-èdè ti LDC àti "àṣàjọ ìwé" fún ọ̀rọ̀-èdè, gẹ́gẹ́ bí àwòkọ́ṣe lórí àwọn àṣeyọrí nínú ìṣẹ̀dá ohun àmúlò, pẹ̀lú ojúṣe LDC nínú akitiyan láti mú kí àkójọpọ̀ ọ̀rọ̀-èdè fún gbogbo àwùjọ ajọlèdè wà ní àrọ́wọ́tó. ti awọn ede ti ko ni idaniloju. Iwe yii ṣe ayẹwo corpora LDC ti o yẹ ati awoṣe "idii ede" awoṣe apẹẹrẹ bi awọn aṣeyọri ni ẹda awọn orisun, pẹlu ikopapọ Consortium ninu awọn ipa ti o ni ilosiwaju si data fun gbogbo agbegbe agbegbe.

1. Introduction

This paper introduces the Linguistic Data Consortium (LDC), an open consortium of universities, libraries, corporations and government research laboratories hosted at the University of Pennsylvania USA and describes how it fulfills its mission to develop and broadly share language resources. LDC's online catalog is a rich, curated repository of multilingual speech, text, video and lexical data sets that includes publications of interest to underserved language communities. The Consortium also works with like-mended global sister organizations and networks to advance language-related research, education and technology development in the world's languages.

2. LDC: Founding, Mission and Operation

LDC was founded in 1992 to address the critical data shortage then facing language technology research and development on the principle that broad access to data drives innovation. Its mission is to support language-related education, research and technology development by creating and sharing linguistic resources, such as data, tools and standards. From its primary role as a repository and distribution point for language resources, the Consortium has grown into an organization that creates and distributes a wide array of language resources to the global community and supports sponsored research programs and language-based technology evaluations by providing resources and contributing organizational expertise.

The Consortium is a mutual aid society. Researchers contribute data sets to the LDC Catalog and as a result, their work gains visibility and community recognition and inspires other research. Members and data licensees contribute fees and in return receive ongoing rights to a variety of resources; those fees are typically a fraction of the cost of data development. Sponsors contribute funding that results in resource creation, infrastructure, innovation, cost sharing and resource dissemination to the community.

3. The LDC Catalog

3.1 Sharing Data in the World's Languages

The LDC Catalog is a growing digital archive of over 800 holdings that for more than two decades has served as one of the world's major language resource repositories. As the first and most active language resource data center, LDC established or adopted many of the publication, archiving and curation practices that related research communities follow today. Originally seeded by data contributions of significant corpora, the catalog continues to be augmented by data sets developed by LDC and by donations from researchers worldwide. As of this writing, LDC has distributed close to 200,000 copies of its resources in over 90 languages to roughly 6000 distinct organizations in more than 100 countries. Over 10,000 unique papers citing LDC data have been identified, attesting to the repository's overall research impact.

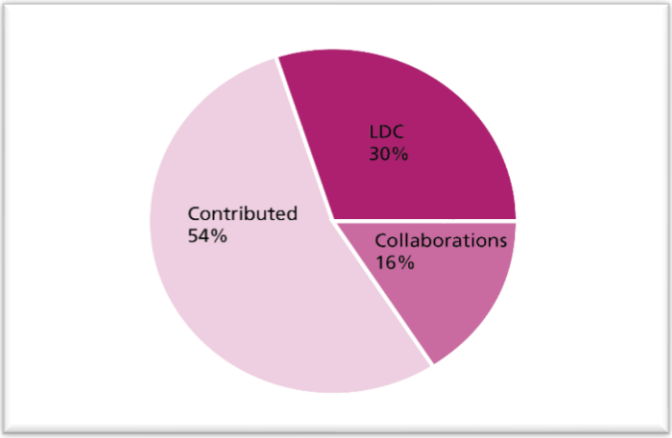

Figure 1: The LDC Catalog is a community resource.

3.2 Curating Language Resources

The catalog has also been recognized as a trustworthy data repository under the CoreTrustSeal certification established by the ISCU World Data System and the Data Seal of Approval. [1] This means that the Catalog meets high standards for data access, rights management, curation, data integrity and authenticity, archival storage and security. The Catalog also consistently receives the highest (five-star) rating for compliance with the Open Language Archives Community (OLAC) metadata standard, an extension of the Dublin Core standard designed for language resources.[2]

LDC's curation workflow includes data review upon submission, a battery of quality checks, metadata creation and documentation development. The data is then prepared for delivery, usually via web download or on media for larger corpora. All data distributed through the catalog is archived in a logical data tree subject to a specialized back-up system from which it can be migrated to new formats, platforms and storage media as required by best practices in the digital preservation community.

LDC's licenses are compatible with the community's customary uses as well as with intellectual property, human subjects and privacy concerns. These include tribal rights in community languages, recently reaffirmed in revised US human subjects regulations.

4. Language Resources Overview

LDC develops and publishes resources in a growing number of languages referred to under several terms: indigenous languages, minority languages, endangered languages and low resource languages. Whatever the name, such languages pose challenges to researchers. Human language technology development relies on digital resources, such as lexicons, grammars, monolingual and parallel corpora, morphological analyzers, taggers and segmenters. For some languages, the source data is scarce; for others the structure of the language itself affects the development of technology-related resources. Below is an overview of some LDC data sets and research noting solutions to language-specific issues.

4.1 West African Languages

Among the research challenges presented by West African languages are complex phonology and morphology (Bantu), verb serialization (Kwa), complex pronoun systems (Yoruba) and the absence of established writing systems (many). LDC data sets in the Manding languages, Yoruba and Dschang and Ngomba (Bantu) illustrate creative and flexible solutions to language challenges.

Grassfields Bantu Fieldwork: Dschang and Ngomba Tone Paradigms. Tonological and phonetic description of tone paradigms.

Global Yoruba Lexical Database. Diaspora dialects included to capture the language's global impact: Nigeria and Benin to the Caribbean and islands along the southeastern United States coast.

Manding lexicons (Bamanankan, Maninkakan, Mawukakan). Bidirectional English and French glosses to accommodate speakers in a francophone context.

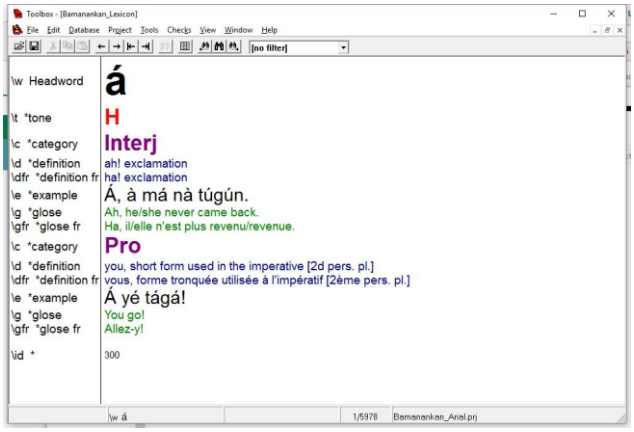

Figure 2: Entry from Bamanankan Lexicon (LDC2016L01) in Toolkit interface

4.2 Fieldwork

Some recent approaches in fieldwork documenting endangered languages incorporate simple technologies like handheld recorders and smartphones to allow large numbers of community members to capture speech for re-speaking, transcription and translation. LDC contributed to two such studies in **Papua New Guinea** and **Brazil** funded by the National Science Foundation (BCS-0951651, IIS-0964556).

Malto is a Dravidian language spoken in northeastern India and Bangladesh by people called the Parahiyas in villages or hamlets located on hilly tracts and in the lowlands. The **Malto Speech and Transcripts** corpus contains audio data from speakers who share their life stories, local rituals from

[1] https://www.coretrustseal.org/.

[2] http://www.language-archives.org/.

festivals to funerals, and the oral histories and rich folklore of their community.

4.3 Language Packs

LDC has developed « language packs » for low resource languages in two US government-funded projects, REFLEX and LORELEI. The idea behind the language pack is to construct a core set of language resources and tools that can be deployed for multiple purposes, among them, language documentation and preservation, basic technology development and situational awareness, e.g., natural and humanitarian disasters. (Simpson, et al., 2008 ; Strassel and Tracey, 2016).

Figure 3: Collaborative transcription in Papua New Guinea (Courtesy: Steven Bird)

Language packs consist of monolingual text, parallel text, several types of annotation, tools for text processing, segmentation and entity tagging, as well as lexicons and grammatical sketches. Languages covered include **Akan (Twi), Amazigh, Amharic, Ilocano, Kinyarwanda, Odia, Oromo, Sinhala, Tigrinya, Uighur, Uzbek, Wolof and Zulu.** It is expected that packs for roughly 20 languages will be released into the LDC Catalog beginning in 2020.

5. Research Collaborations in Indigenous Languages

LDC is an active participant in related consortia and groups whose aim is to advance the ways in which resources are developed and distributed. These include initiatives for standardizing specifications and best practices and for developing new architecture to support language resource delivery. Collaborations involving indigenous languages are among these.

5.1 Community Projects

In the National Science Foundation's **AARDVARC project** (Automatically Annotated Repository of Digital Audio and Video Resources Community), [3] LDC engaged with an interdisciplinary community of linguists, anthropologists and computer scientists to discuss and develop standards around formats, access and use of resources in endangered and low resource languages.

Similarly, in **E-MELD** (Electronic Metastructure for Endangered Languages Data), [4] LDC participated in the effort to develop consensus on documenting endangered languages and fostering collaboration among digital archives.

5.2 Languages of the Americas

LDC promotes resource development in the Americas in a variety of ways. These include advice and technical assistance for specific collections, among them, Nahuatl and Mixtec. Recently, the Consortium convened two workshops in 2018 exploring hemispheric collaboration and language resource development.

The **Planning Workshop on Data Archives and Languages of the Americas**[5] was held in Philadelphia with support from the University of Pennsylvania. Experts managing linguistic data archives and resource centers met to discuss challenges, needs and opportunities for promoting and extending collaboration in the Americas.

The **International Workshop on Data Intensive Research on Languages in the Americas,**[6] also supported by the Penn Global initiative at the University of Pennsylvania, took place in Mexico City. Linguists and scientists from Mexico, Brazil, Chile, Argentina and the United States presented their work on Chuj, Yucateco, Huasteco, Nahuatl, Wixarika, Southern Cone languages, Mexican/American Spanish and Brazilian Portuguese.

These collaborations have provided the beginnings of a strong regional community and the basis for future work.

Figure 4: LDC's global network of contributors and collaborators

6. Innovation in Language Resource Development

Despite the large volumes of linguistic data created by current methods, supply continues to lag far behind demand. This is due in part to the application of a finite resource to a problem that is effectively infinite or at least several orders of magnitude larger.

LDC has recently begun to address this problem by identifying renewable sources of the time and intellectual investment required to document the world's languages, especially for the purposes of human language technology development. The experience of social media platforms, grass roots efforts such as Librivox, which creates

[3] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1519887.

[4] http://emeld.org/.

[5] https://www.ldc.upenn.edu/communications/workshops/penn-urf-sas-workshop.

[6] https://www.ldc.upenn.edu/communications/workshops/penn-gef-americas-workshop.

audiobooks from out-of-copyright texts, and especially citizen science platforms, demonstrates that the human drive for challenge, advancement, entertainment and the opportunity to contribute to one's own betterment and that of one's local community and the broader society are effectively boundless. For example, nearly two million contributors to the Zooniverse citizen science portal have submitted more than 250 million judgments that are used by researchers in astronomy, biology and other fields.

LDC's *LanguageARC* presents language resource projects to potential Citizen Linguists. Each includes multiple tasks that require a simple judgment repeated over multiple items. For example, one project might seek to document the state of a number of indigenous languages of South Africa through surveys that document the point in children's development at which they acquire the words for culturally significant objects. Another might elicit local terms via picture or silent video description. Still others might elicit re-speaking or translations as a way to reveal the grammatical features of a language. LanguageARC supports any task in which contributors are shown a text or images or are played audio or video clips and asked to respond to instructions that are either specific to the task or that vary with each item by speaking, entering a text response or selecting one or more items from a multiple choice list.

To attract and support a community of contributors each LanguageARC project has a title, call to action, image, pitch, picture, partner badges, description of the research team and discussion forums to support community building. Tasks similarly have a title, calls to action and images but also include tutorials, reference guides and their own discussion forum.

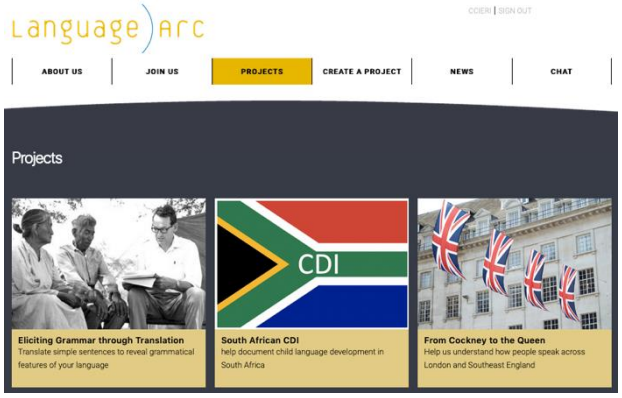

Figure 5: The LanguageARC Citizen Linguist portal

7. Conclusion

Access is a crucial theme of the 2019 International Year of Indigenous Languages – access to education, information and knowledge for indigenous peoples in their home languages. Means to that end include the availability of data collections that in turn can be used to develop language technologies for indigenous language communities. LDC's founding principle that broad access to data drives knowledge and research resonates with that theme. The Consortium is committed to developing and sharing resources in all languages for all language communities in ways that ensure meaningful access, advance language vitality and promote preservation.

8. Bibliographical References

Bamanankan Lexicon. (2016). Distributed via LDC, LDC2016L01, ISLRN 830-816-122-814-4.

CoreTrustSeal. https://www.coretrustseal.org/. Accessed 25 November 2019.

EMELD. http://emeld.org/, Accessed 25 November 2019.

Federal Policy for the Protection of Human Subjects. 82 Fed. Reg. 7149 (Jan. 19, 2017).

Global Yoruba Lexical Database v 1.0 (2008). Distributed via LDC, LDC2008L03, ISLRN 973-344-578-516-8.

Grassfields Bantu Fieldwork: Dschang Lexicon. (2003). Distributed via LDC, LDC2003L01, ISLRN 880-081-036-797-6.

Grassfields Bantu Fieldwork: Dschang Tone Paradigms. (2003). Distributed via LDC, LDC2003S02, ISLRN 973-117-906-652-9.

Grassfields Bantu Fieldwork: Ngomba Tone Paradigms. (2001). Distributed via LDC, LDC2001S16, ISLRN 147-689-240-962-1.

International Workshop on Data Intensive Research on Languages of the Americas. https://www.ldc.upenn.edu/communications/workshops/penn-gef-americas-workshop. Accessed 25 November 2019.

Malto Speech and Transcripts. (2012). Distributed via LDC, LDC2012S04, ISLRN 841-757-472-203-8.

Maninkakan Lexicon. (2013). Distributed via LDC, LDC2013L01, ISLRN 573-342-913-646-6.

Mawukakan Lexicon. (2005). Distributed via LDC, LDC2005L01, ISLRN 592-356-503-307-6.

National Science Foundation. Collaborative Research: Automatically Annotated Repository of Digital Video and Audio Resources Community (AARDVARC). https://www.nsf.gov/awardsearch/showAward?AWD_ID=1519887. Accessed 25 November 2019.

OLAC: Open Language Archives Community. http://www.language-archives.org/. Accessed 25 November 2019.

Planning Workshop on Data Archives and Languages of the Americas. https://www.ldc.upenn.edu/communications/workshops/penn-urf-sas-workshop. Accessed 25 November 2019.

Simpson, H., et al. (2008). Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), W10, Collaboration: interoperability between people in the creation of language resources for less-resourced languages, pages 7-11, Marrakesh, Morocco, May. European Language Resource Association (ELRA).

Strassel, S. and Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3273-3280, Reykjavik, Iceland, May. European Language Resource Association (ELRA).