Trends in HLT Research: A Survey of LDC's Data Scholarship Program

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

Abstract

Since its inception in 2010, the Linguistic Data Consortium's data scholarship program has awarded no cost grants in data to 64 recipients from 24 countries. A survey of the twelve cycles to date – two awards each in the Fall and Spring semesters from Fall 2010 through Spring 2016 – yields an interesting view into graduate program research trends in human language technology and related fields and the particular data sets deemed important to support that research. The survey also reveals regions in which such activity appears to be on a rise, including in Arabic-speaking regions and portions of the Americas.

Keywords: language resources, data, graduate studies, human language technology

1. Introduction

The Linguistic Data Consortium's (LDC) Data Scholarship program [1] was established in 2010 and formalized the Consortium's long-standing principle that no one with a bona fide research agenda and a genuine inability to contribute should go without data. Before the grant program, LDC routinely handled requests from (principally) graduate students who needed particular data sets for their thesis or dissertation work and whose institutions lacked the financial resources to acquire them. The Consortium developed a formal program subsidized by its members so that, consistent with LDC's mission, all in the community would have an opportunity to seek such assistance. The requirements are not onerous. Applicants must submit a data use statement that describes the research plan, use of data and method for determining success as well as a letter of support from their advisor that includes information about the probability of success and asserts inability to contribute. The program is widely advertised on LDC's web pages, social media platforms, in the monthly newsletter, at conferences and through other LDC networks. Since 2010, 64 recipients from 26 countries have received 110 corpora valued at over USD $175,000. There have been 100 scholarship applicants overall, yielding a success rate of 64%.

This paper describes the data scholarship program, its requirements and evaluation criteria, followed by a survey of applicants, requested data sets, research areas, and the countries represented. That quantitative data is then analyzed for signs of research trends and the geographic areas in which the work is being conducted.

2. Data Scholarship Program Details

Data scholarships are offered semiannually during the fall and spring semesters and are available to students pursuing undergraduate or graduate studies in an accredited college or university. They are not restricted to any particular field of study but, in keeping with LDC licenses, projects must deal with linguistic aspects of the resources.

The application has two components. A **Data Use Proposal** must include: the name of the database(s) requested, a brief description of the research project, a description of how the data will be used and how success will be measured. The **Letter of Support** from the applicant's thesis/dissertation advisor or department chair must verify the student's need for data and confirm that the department or university lacks funding to pay the applicable nonmember license fee or to join the Consortium. The letter must also express the advisor's confidence that the project will be successful.

Applications that are incomplete, that do not meet the requirements stated above or are received after the submission deadline are rejected. Applicants who demonstrate the following are more likely to have their request more highly rated:

An understanding of the database(s) requested. For example, if a proposal aims to develop an entity tagging technology, then the requested corpus should be entity-tagged or the proposal must make a provision for adding such tags.

An evaluation methodology appropriate to the type of research proposed. For example, for research in speech recognition, a proposal which plans to use a database, evaluation protocol and scorer already used in recognized evaluation campaigns would be rated higher than one that did not.

A research methodology appropriate to the student's field. For example, proposals that adopt an accepted methodology or else motivate an alternative methodology will be better rated than those that simply adopt a new methodology without justification.

Appropriate planning. For example, if a proposal aims to process a very large corpus in a short amount of time, the student should mention how necessary computer resources will be deployed.

These considerations are quite important and have been refined to reflect our experiences over the life of the

[1] https://www.ldc.upenn.edu/language-resources/data/data-scholarships

program thus far. We have observed that applicants do not always understand what a corpus contains or how it should be used for a particular task, nor is there universal understanding about how technology systems should be trained and tested. Asking applicants to illuminate these points allows the reviewers to better gauge a research plan's feasibility.

Those who cannot provide that information typically fail. We think that some failures are the result of applicants simply not making the necessary effort to develop a good application. In certain instances, advisors are not enthusiastic about, or do not address, the probable success of the proposed work. We also see in particular situations that the academic program may fail to provide guidance for designing effective experiments to test theoretical principles.

3. Award History

Following is a survey of data scholarship awards to date by country, by field of study and by corpus type.

3.1 By Country

Of the 64 awards to date, the largest number of recipients were in the United States [15], followed by India [10], Mexico [4] and China [4]. The others are scattered across the globe, with clusters in Arabic-speaking countries (Algeria, Egypt, Jordan, Lebanon), the Americas (Argentina, Brazil, plus Mexico as indicated above) and Asia (Indonesia, Malaysia and China). Table 1 shows the number of award recipients per country.

Country	Recipients
Algeria	1
Argentina	1
Brazil	2
China	4
Egypt	2
Finland	1
Greece	1
Iceland	1
India*	10
Indonesia	1
Iran	2
Ireland	1
Italy	1
Jamaica	1
Japan	3
Jordan	1
Lebanon	2
Malaysia	2
Mexico*	4
New Zealand	1
Switzerland	1
Tunisia	1
Turkey	1
UK	3
Ukraine	1
USA	15

Table 1: LDC Data Scholarship Recipients by Country (*Some awards were made to a research group, but are counted here as a single recipient)

On the one hand, it may not be surprising that there are institutions around the world that are underfunded. This particular survey suggests, however, that there are certain regions where the need for basic linguistic resources may be more acute. It may also demonstrate the influence of available resources for research. For instance, a number of Arabic language data sets have been developed over the last decade or so, and their availability may have helped to inspire further work in the Arabic language as well as among researchers in Arabic speaking areas of the world. The same may be true for research in Chinese. The scholarship applicants from the Americas are from diverse institutions, indicating that HLT research in the region is spreading and that a growing range of other disciplines are beginning to embrace big data approaches.

In the case of the United States, the scholarship award history suggests that it is not as unusual as one might think to find underfunded US computer science, linguistics and engineering departments resulting in the large number of applications.

3.2 By Field of Study

The majority of scholarship candidates are graduate students in computer science, electrical engineering, informatics and related fields. There have been a handful of applicants from linguistics, applied linguistics and the social sciences (e.g., psychology).[2] The fields in which the awards have been used are consistent with that candidate profile.

Of the resources awarded to date, the vast majority of them were to be used for speaker recognition research. Information extraction and speech recognition were the next most popular fields, but they each constituted less than half of the number of awards attributable to the leader.

Out of 64 award categories (some recipients were working in more than one related field), the leaders were: speaker recognition and diarization (20), information extraction (7), speech recognition (5), data mining (3) and anaphora resolution (2), followed by single awards across various disciplines. A full list of award categories appears in Table 2.

[2] We have made additional efforts to attract candidates from linguistics and related fields through in-person contacts at relevant conferences, including NWAV (New Ways of Analyzing Variation) and the Linguistic Society of America's annual meeting.

Field of Study	Awards
anaphora resolution	2
acoustic modeling	1
code-switching, speech	1
data mining	3
diacritization (Arabic)	1
disambiguation	1
document retrieval	1
emotional speech	1
handwriting recognition	1
induction	1
information extraction	7
eye movement tracking	1
language identification	2
machine learning	3
parsing	2
prosody	1
psycholinguistics	1
signal processing	1
semantics	1
sentiment analysis	1
speaker recognition, diarization	20
speech recognition	5
spoken term detection	1
summarization	1
syntax	1
tagging	1
topic detection	1
voice activated system	1

Table 2: LDC Data Scholarship Awards by Field

We think that one can correlate the concentration in speaker and speech recognition research among graduate students to current market forces. The rise of speech-dependent customer service centers and mobile phone applications, as well as the growing need for better speech synthesis systems (e.g., in-car navigation tools) is reflected in industry reports as well as at conferences featuring industry solutions, like SpeechTEK. [3] Over the last few years, analysts have made bold predictions about the expected growth of the speech technology market. For instance, one source stated that the global voice recognition market would exceed US$100 billion by 2017, driven by voice biometrics, primarily for mobile devices.[4] Others expect the speech analytics market (described as "audio mining") to be valued at over US$1 billion by 2019.[5] To the extent that graduate studies are geared to economic developments with the potential to yield job prospects for graduates or research funding, the high interest in speaker and speech recognition makes sense.

3.3 By Corpus Type

By far, data sets that represent evaluation training, development and/or test material are the most popular selections from among data scholarship candidates and winners. The US National Institute of Standards and Technology (NIST) evaluation corpora for speaker recognition (SRE), are among the most requested resources, followed by the ACE corpora for content extraction. The collections underlying SRE evaluations, such as the CALLHOME and Switchboard telephone studies, are also highly valued, as are other benchmark data sets that include HUB4 broadcast news and transcripts, the TIMIT series, continuous speech recognition (CSR), TIDIGITS and YOHO. For text analysis, LDC's Gigaword corpora, Topic Detection and Tracking (TDT) data sets and various Treebanks are all common choices. For some fields, unique data sets in the LDC Catalog are required, such as for emotional speech (Emotional Prosody) and handwriting recognition (MADCAT Training Data).

4. Challenges

The Data Scholarship program has not been without its administrative challenges. Principal among these is the tension between the desire to support young scholars and the need to remain good stewards of Consortium funds. The program is supported solely by members' annual fees that pay the salaries of the review committee and the costs of replicating and delivering data. The committee must make difficult decisions concerning whom to award and these are based only on an estimation of the probability that the research program will succeed and contribute knowledge to the field.

A second challenge arises from the wide range of scientific disciplines represented among the applicants, While LDC never intended that the review committee be expert in every appropriate field, we nonetheless must recognize that, for example, it is reasonable to demand a metrics-driven evaluation for some research projects but not others where the metrics, gold standard data, scorer or even the concept may be absent.

The truly international nature of the applicant pool makes this an exciting program, but brings additional challenges. The authors and review committee members have experience as reviewers for other funding bodies in the United States and abroad. In some of those panels, knowledge of a researcher's previous work or mentors may provide useful background. However, within the Data Scholarship program, applicants are first stage researchers often from unfamiliar groups, removing a source of information.

5. Successes

Despite the challenges involved in the LDC Data Scholarship program, we believe it makes a valuable

[3]SpeechTEK 2015, http://www.speechtek.com/2015/AdvanceProgram.aspx.

[4]Global Voice Recognition Market Forecast to Reach $113 Billion by 2017, http://www.speechtechmag.com/Articles/News/Speech-Technology-News-Features/Global-Voice-Recognition-Market-Forecasted-to-Reach-$113-Billion-by-2017--96508.aspx

[5] Speech Analytics Market worth $1.33 Billion by 2019, http://www.marketsandmarkets.com/PressReleases/speech-analytics.asp.

contribution to multiple fields. A goal for the program has been to aid new developments in language-related research and technology, and based on feedback from award recipients, we think that goal has been met.[6]

For instance, most reported that they used the data as they intended and received the results they expected. Three students have graduated from their programs and two more expect to graduate in 2016. There have been at least six published papers based on data received in the program. (E.g., Guven, 2012; Harrat, et al., 2013; George, et al., 2015). Most awardees described the data they received as vital to their work. In one case, data awarded through the program was used to build a state-of-the-art speaker recognition system (AMRITATCS) that the awardee and his colleagues submitted to The Speakers in the Wild (SITW) Speaker Recognition Challenge hosted by SRI International.[7]

There were some who found that they could not use the data, for instance, because they expected it to contain something it did not, the data set was too small, or their dissertation topic changed. Overall, however, as indicated above, students report positive experiences in the program.

We think that the data scholarship program has helped potential new entrants to the field by giving them the experience of working with data as they will be expected to in their careers. It is clear to us, however, that metric-driven evaluation has not spread throughout the field. This is something we as a community should consider addressing to ensure that new entrants are prepared to make meaningful contributions and further progress.

6. Other Programs

We are unaware of any other program serving students in the community like LDC's data scholarship program. Among data centers, ELRA offers some of its language resources data at no cost and offers internships. Similarly, Gengo Shigen Kyokai (Japan) provides its data to users at no cost, charging handling fees only. Several data sets in LDC's catalog are likewise available to non-members at no cost. However, the data scholarship program covers all of LDC's 600+ holdings without restriction.

LDC-IL issues occasional applications for short term projects which may attract student candidates.[8] The US National Science Foundation (NSF) provides infrastructure awards, planning grants and travel support, all of which benefit students, but none of which involve providing students specifically with data, though students may use funds for that among other purposes. CLARIN ERIC (European Union) has a Mobility Grant scheme that allows researchers and technical staff working at an organization in a CLARIN member country to travel to a CLARIN center in another member country for a week to train or collaborate on matters affecting CLARIN..

7. Future Directions

The program works well in its current form, but might not be sustainable indefinitely. One option under consideration is supplemental funding for the awards, thereby lessening the burden on Consortium members. We also continue to explore ways to reduce the costs of data distribution. We now provide more resources via electronic transfer, from LDC, the cloud or grid. We expect market costs for bandwidth to decline, along with storage costs; but those expenses are still significant for data of a certain size. Current innovations in LDC's business system allowing e-signature on licenses and the ability to license data online are efficient and user-friendly. Nevertheless, we still must have the capability to answer questions about data sets and troubleshoot user issues.

We welcome comments and suggestions from the community about future directions.

8. Conclusion

The LDC Data Scholarship program began in 2010 as a way to formalize the Consortium's support of numerous research communities. Twice each year students apply by submitting a Data Use Statement and Letter of Support from their advisor. A review committee at LDC evaluates the proposal along a number of factors that, we believe, equate to probability of success and to a contribution to the field. More than one quarter of the applicants come from US universities. India, China and Mexico are also well represented. Applications from the Americas, the Arabic speaking countries and Asia are growing.

The most well represented fields are Speaker Recognition and Information Extraction. The most popular corpora include technology evaluation data sets, Treebanks, Gigaword text corpora and telephone speech studies. Challenges derive from the highly international nature of an early career applicant pool. However, a number of successes and even a few testimonials justify the expense of the program.

9. Acknowledgements

Thanks are due to the members of the Linguistic Data Consortium who subsidize the Data Scholarship program through their membership fees.

10. Bibliographical References

Central Institute of Indian Languages, Circular, http://www.ldcil.org/download%5CSTGOP2012CIRCULAR.pdf

George, Kuruvachan K., Santosh, Kumar C., Ramachandran, K.I., Ashish, Panda. (2015). Cosine distance features for robust speaker verification. In *Proceedings of the 16th Annual Conference of the*

[6] The information that follows is based on recipient responses to an informal survey by LDC asking participants if they had published a paper, graduated from their program and received the results they expected from the data.

[7] http://www.speech.sri.com/projects/sitw/

[8] Central Institute of Indian Languages, Circular, http://www.ldcil.org/download%5CSTGOP2012CIRCULAR.pdf

Internal Speech Communication Association. Dresden, Germany.

Global Voice Recognition Market Forecast to Reach $113 Billion by 2017, http://www.speechtechmag.com/Articles/News/Speech-Technology-News-Features/Global-Voice-Recognition-Market-Forecasted-to-Reach-$113-Billion-by-2017--96508.aspx.

Guven, Erhan. (2012). Robust classification of emotion in human speech using spectrogram features. Doctoral Dissertation. George Washington University.

Harrat, Salima, Abbas, Mourad, Meftouh, Karima, Smaili, Kamel. (2013). Diacritics restoration for Arabic dialect texts. In Proceedings *of the 14th Annual Conference of the Internal Speech Communication Association.* Lyon, France.

LDC Data Scholarships, https://www.ldc.upenn.edu/language-resources/data/data-scholarships.

Speech Analytics Market worth $1.33 Billion by 2019, http://www.marketsandmarkets.com/PressReleases/speech-analytics.asp.

SpeechTEK 2015, http://www.speechtek.com/2015/AdvanceProgram.asp.

The Speakers in the Wild (SITW) Speaker Recognition Challenge, http://www.speech.sri.com/projects/sitw/.

11. Language Resource References

Campbell, Joseph, and Alan Higgins. (1994). YOHO Speaker Verification, distributed via Linguistic Data Consortium, LDC94S16, ISLRN 125-762-148-524-1.

Lee, David, et al. (2012). MADCAT Phase 1 Training Set, distributed via Linguistic Data Consortium, LDC2012T15, ISLRN 1-58563-623-1.

Leonard, R. Gary, and George Doddington. (1993). TIDIGITS, distributed via Linguistic Data Consortium, LDC93S10, ISLRN 177-353-807-744-3.

Liberman, Mark, et al. (2002). Emotional Prosody Speech and Transcripts, distributed via Linguistic Data Consortium, LDC2002S28, ISLRN 191-383-337-125-7.