LDC Language Resource Database: Building a Bibliographic Database

Eleftheria Ahtaridis, Christopher Cieri, Denise DiPersio

Linguistic Data Consortium, University of Pennsylvania

3600 Market Street, Suite 810, Philadelphia, PA 19104 USA

{ilya, ccieri, dipersio}@ldc.upenn.edu

Abstract

The Linguistic Data Consortium (LDC) creates and provides language resources (LRs) including data, tools and specifications. In order to assess the impact of these LRs and to support both LR users and authors, LDC is collecting metadata about and URLs for research papers that introduce, describe, critique, extend or rely upon LDC LRs. Current collection efforts focus on papers published in journals and conference proceedings that are available online. To date, nearly 300, or over half of the LRs LDC distributes have been searched for extensively and almost 8000 research papers about these LRs have been documented. This paper discusses the issues with collecting references and includes preliminary analysis of those results. The remaining goals of the project are also outlined.

Keywords: bibliographies, digital libraries, metadata

1. Background

The Linguistic Data Consortium (LDC) is an open, not-for-profit consortium of universities, corporations, and government research organizations that creates and provides language resources (LRs) including data and tools and specifications. LDC currently distributes over 500 LRs and adds approximately 30 new ones to its Catalog[1] each year. In addition to those in the catalog, LDC develops and distributes LRs for sponsored projects and common task evaluations. Including both types, LDC has distributed over 84000 copies of more than 1300 titles to 3100 organizations in 70 countries. LDC's LRs are used by researchers, technology developers, teachers and others engaged in a variety of language-related and human language technology tasks, including language and speaker recognition, information retrieval and extraction, machine translation, language learning and natural language processing.

In order to inform the research communities it supports about its resources and its work, LDC staff have written, presented and published over 150 papers and book chapters. Similarly, LDC members and data licensees have produced thousands of papers describing their use of these resources, the technologies they have created and their research findings. Such papers also routinely report any further conditioning of the data sets, derived LRs, issues encountered and solutions adopted. Unfortunately, there is no single repository where one can find all of the academic papers associated with an LR and no standard way to link LRs to each other or to associated papers. As a result, each user must independently face the problem of learning about an LR. This leads inevitably to duplication of effort and accidental variation in researchers' knowledge about LRs. Of course, the problem of access to information about LRs has been long acknowledged. Bird and Simons (2000) recognized the need to catalog "advice" when they were planning the Open Language Archives Community (OLAC), which includes hundreds of academic papers. However, there are currently no links between papers and LRs in OLAC-related or any of the other LR metadata repositories known to the authors. This paper describes LDC efforts to address these and related problems.

2. Motivation

In 2009, LDC began to collect research papers that introduce, describe, extend or rely upon LRs in the LDC Catalog (Cieri, 2009). The motivation was many-fold: (1) to assess the academic impact of LRs, (2) to support the claim that papers about LRs are also LRs, (3) to support LRs users by allowing them to easily find all papers related to an LR, (4) to support LR authors by promoting their efforts and allowing them to track all relevant feedback, and (5) to help LR authors assess their impact on research. Without wishing to enter the debate about citation metrics, the impact of an LR is some function of the number of others using it and the nature of those uses. Collecting metadata about all papers mentioning an LR is the first step in understanding that impact. A resource that

[1] http://www.ldc.upenn.edu/Catalog/

hosts such metadata may enhance the use of the LR and of course, the content of such papers is useful for subsequent users of the LR and for providing feedback to the creator. The goal then is to create a database which would eventually be integrated into the LDC Catalog with bidirectional links between the LRs and papers mentioning them. Since LDC supports many research communities, we expect this bibliography to be accessed by a range of individuals with varied interests in the data. The database would serve multiple user types: end users of LDC's data, LR creators and paper authors. Someone searching for a particular LR could readily identify research that cites the LR and answers questions including the motivation for why the LR was created, how it was developed and used, and any known limitations of the LR or subsequent enhancements developed. Additionally, paper authors could gain greater exposure for their research by adding citations of their work to the database.

3. Methodology

3.1 Selection

At the commencement of the project, LDC assembled a small revolving team of student research assistants to begin building the database. The students came from diverse academic backgrounds at both the undergraduate and graduate levels. All had some experience with web-based social science research. Research assistants were provided with basic background on LDC and some training on how to search for an LR and which papers to include. Research assistants began populating the database with papers describing the earliest LDC LRs published. Not surprisingly, searches of older benchmark corpora, such as *TIMIT* (Garofolo et al, 1993a) resulted in hundreds of papers that required several months to read, evaluate and add to the bibliography. To understand the variation in the impact of LRs, to give the team a sense of accomplishment, and to avoid duplication of effort, the team was divided into two and half were directed to begin working on the most recent published LRs. After a few months, the selection strategy was modified again to identify representative major LR types – broadcast news speech and transcripts, conversational telephone speech and transcripts, multilingual collections supporting language identification, multi-speaker collections supporting speaker recognition, treebanks, news text, lexicons, translations and multiple translation corpora – which were then selected from across the entire LDC catalog. As a result, the range of LR types currently in the bibliography is nearly representative of that in the LDC Catalog.

The database includes scholarly papers with various relationships to the LR cited: those that introduce a new LR, those that describe an existing LR, those that extend an LR, and those that rely upon an LR. The last category, papers that relied upon an LR, constitutes the bulk of the papers archived to date. Those papers that only mentioned an LR in passing and did not work significantly with it were not included.

3.2 Search Process

Searching began with EBSCO MegaFILE®[1], a journal archive available to LDC through the University of Pennsylvania Library. EBSCO MegaFILE was selected since it provides comprehensive, multi-disciplinary coverage of scholarly journal articles. Research assistants searched in EBSCO MegaFILE for each LR's title, including variants, and LDC catalog number (e.g., LDC93S1). The yield from that search was low; only about 300 journal articles, or less than one paper for each LR in the catalog. Moreover, EBSCO MegaFILE does not return results from conference proceedings where many of the papers dealing with LDC LRs appear. To locate conference papers and additional journals not covered in EBSCO, the search next turned to the web.

The team used two journal search engines, CiteSeerX®[2] and Google Scholar®[3] for the next phase of the search. CiteSeerX was chosen for its focus on scholarly articles in computer science. Google Scholar was utilized for its extensive indexing of journals and conference proceedings from many scholarly publishers. Google Scholar consistently yielded the most potential results and was the main search engine used; CiteSeerX yielded more 'on-target' results that contained less article repetition. That is, CiteSeerX provided greater precision while Google Scholar provided greater recall.

[1] http://search.ebscohost.com/

[2] http://citeseerx.ist.psu.edu/

[3] http://scholar.google.com/

3.3 Format

As papers were located, they were formatted and stored using EndNote®[1] bibliographic software. EndNote was chosen since it allows multiple users to collect and organize references as well as to create a searchable bibliography. Work began with a free web version of EndNote; an upgraded licensed version was later acquired for better data management. In particular, the licensed version has enhanced search and reporting features and allows for an unlimited number of references. At minimum, each bibliographic record includes: Author(s), Title, Year (of Publication or Conference), Journal Name or Conference Name, Abstract, URL, and LR(s) used. The LR used is indicated in the 'Notes' field in each EndNote record. A small number of records contain additional information fields including digital object identifier (DOI).

4. Problems encountered

The most common problem encountered was uncertainty about which LR was actually being used in the research the papers described. This uncertainty was due primarily to missing or inadequate citations of the LR. For example, consider an article that mentions using a *Switchboard* corpus, but does not specify which. As shown in Table 1, the LDC Catalog has ten corpora with Switchboard in the name including the original, *Switchboard-1 Release 2* (Godfrey & Holliman, 1997), cellular versions, and transcripts.

Catalog Number	Corpus Name
LDC93S8	Switchboard Credit Card
LDC97S62	Switchboard-1 Release 2
LDC98S75	Switchboard-2 Phase I
LDC99S79	Switchboard-2 Phase II
LDC2001S13	Switchboard Cellular Part 1 Audio
LDC2001S15	Switchboard Cellular Part 1 Transcribed Audio
LDC2001T14	Switchboard Cellular Part 1 Transcription
LDC2002S06	Switchboard-2 Phase III Audio
LDC2004S07	Switchboard Cellular Part 2 Audio
LDC2009T26	NXT Switchboard Annotations

Table 1: LDC's Switchboard corpora

In such a case, the team reviewed LDC's licensing records for the primary article author. If the author's organization had licensed only one *Switchboard* data set, this would strongly suggest the corpus intended by the reference. The team also checked the date of publication of the paper against the release date of the given LR. In the example above, if the article pre-dated the release of all *Switchboard* corpora but one, this would conclusively indicate the LR cited.

The team encountered many articles for which they could not determine the specific LR used. In those cases, the LR identified in the database is the project 'family' name; so in the current example, just *Switchboard*. This issue unfortunately affects some of the most-cited resources in LDC's corpus catalog including the English Penn Treebanks (Marcus, Santorini, & Marcinkiewicz, 1995 and Marcus et al., 1999).[2] A related problem was that papers often referred to the LR by the data source and not its official name. For instance, many papers referenced a Wall Street Journal (WSJ) corpus; WSJ data is included in several LDC LRs, including *CSR-I* (Garofolo et al., 1993b), *BLLIP* (Charniak et al., 2000), *Penn Treebank* and *TIPSTER* (Harman & Liberman, 1993). Roughly 10% of papers read lack a reference to the specific LR used.

It is also interesting to note that most papers failed to cite LRs in the 'references' section. That is, the LR referenced in the paper was often not treated in the same manner as other research papers. Unfortunately, citations of corpora and other data sets have not yet taken hold in most research communities using LRs. LDC provides citation guidelines for each LR in its Corpus Catalog. The guidelines include the LR author, year of publication, LR name, and Linguistic Data Consortium as the LR publisher.

A related issue is that many papers make no mention of LDC, either in the reference section or in the paper itself. For instance, a Google Scholar search for the LDC LR, WSJCAM0 Cambridge Read News (Robinson, et al.,

[1] http://www.endnote.com/

[2] LDC distributes two versions of the Penn Treebank: Treebank-2 (LDC95T7) and Treebank-3 (LDC99T42)

1995), returns 208 possible results. A search of both "WSJCAM0" and "LDC" yields forty three possible results, and for "WSJCAM0" and "Linguistic Data Consortium" yields just twenty one possible hits. This means that searchers could not simply include LDC in search terms to limit results and focus on those that were likely to be on target. To better understand the issue, consider that a Google Scholar search for LDC LR "Switchboard" returns over 55,000 possible hits since the term "switchboard" is used in many disciplines. Results cannot be limited by restricting the search for those articles that also mentioned LDC since this could exclude many potential articles. Instead, searchers must resort to less effective tactics to focus on relevant papers, such as restricting the subject area search or publication date of the article.

Searches often returned several URLs for a given paper. When possible, the URL chosen was that of a database such as the ACL Anthology, which provides access to the full text of the paper. Linking to personal web pages was avoided since those pages are less likely to be maintained over time. Another challenge was finding full-text URLs for some articles. Google Scholar's search results are often linked to 'reader-pays' journal article archives such as ACM Digital Library [1], IEEE Xplore® [2], and SpringerLink [3] Those sites generally require a subscription or fee for full article access although some information about the article including a summary may be provided at no cost. In such cases, the team first attempted to locate a freely-available version of the paper. If a no-cost version could not be found, articles that could be viewed with the journal subscription services provided by the library of the University of Pennsylvania, LDC's host institution, were included. If the team could not access the full text of the paper, they included the paper in the bibliography only if they could determine from any summary or other information provided that the research used a given LR.

[1] http://dl.acm.org/

[2] http://ieeexplore.ieee.org/Xplore/guesthome.jsp

[3] http://www.springerlink.com/

5. Progress to date

Currently, the papers database contains almost 8000 references which represent an in-depth search of approximately 55% of the LRs in the catalog. Specifically, this includes LDC catalog years 1993 through 1995 and 2008 and 2009, as well as LRs representative of each type LDC distributes (broadcast, lexicons, treebanks, etc.). A small number of corpora not targeted in those years and genres are also archived in the database since a search for one particular LR often returned hits for other LRs that were either used as part of the same research project or that had similar names.

The papers database focuses on published LRs in LDC's Catalog. Although searches often returned papers describing unpublished LRs, those were excluded from the bibliography unless they also involved significant work with a published LR. Unpublished LRs were not searched for specifically since the yield of papers is expected to be lower and concentrated in evaluation workshop proceedings (which are also less likely to be published). As unpublished LRs constitute over half of LDC's LRs, it is expected that LDC's research impact is larger than that captured by the current project.

6. Current Analysis

As almost 300 LRs, over half of LRs in LDC's catalog, have been searched for extensively, some preliminary analysis on results is now possible. The licensed version of EndNote allows for searching within a field and across fields, so the number of times an LR was cited can be counted. One question was whether those LRs that are licensed most frequently are also those that are referenced most frequently in academic papers. To investigate this relationship, the number of licenses for a given LR was compared to the number of times that LR was cited in a research paper.

Table 2 plots the number of licenses[4] for a given LR along the X axis against the number of times that LR was referenced in a research paper along the Y axis. As

[4] The license count includes those LRs distributed for evaluation use and those distributed to LDC's Subscription Members, who receive two copies of each LR.

expected, the more times an LR has been licensed, the more that LR is used, so there is a positive relationship between the two variables (correlation coefficient .633). Of those LRs searched for extensively, an average of 38 papers per LR have been located with a range of 0 to 692 papers per LR.

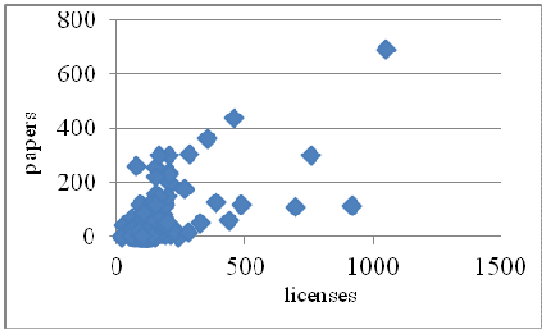

Table 2: License count versus paper counts for LRs

To further investigate the relationship between resources and research, the number of papers for LDC's most licensed LRs, the "Top Ten"[1] , were examined. Table 3 shows (1) the number of licenses for and (2) papers about LDC's most licensed LRs.

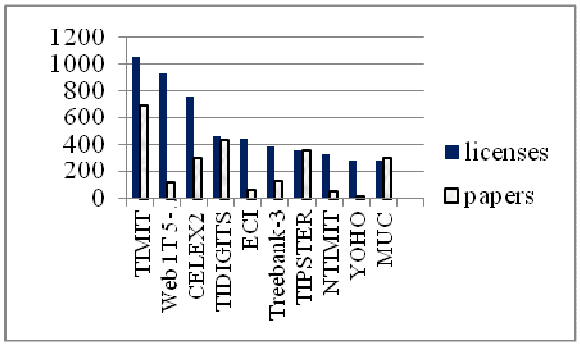

Table 3: License and paper counts for LDC's Top Ten LRs

A few trends emerge – as before, the more times an LR has been licensed, the more likely it is to be used in research. For certain LRs that have been licensed frequently yet do not appear very often in scholarly literature, such as Web 1T 5-gram (Brants & Franz, 2006), and YOHO Speaker Verification (Campell & Higgins, 1994), a few explanations are possible. One possible explanation is that these LRs were simply not mentioned by the paper author. In the case of an LR such as Web 1T 5-gram, it may also be that since this LR is fairly new compared to the other Top Ten LRs[2], research use of this LR is still in its infancy.

On the other hand, an LR such as Message Understanding Conference (MUC) 7 (Chinchor, 2001) has been referenced more than it has been licensed. This data, as well as other benchmark databases such as TIMIT, pre-date the establishment of LDC, and thus, there are more organizations with access to this data than LDC's license count captures.

7. Future Work

The papers database currently resides on LDC's network. Remaining steps include (1) completing the initial bibliographic search for all LDC LRs, (2) extending OLAC or a similar metadata repository to include links among LRs, (3) converting the EndNote database into a web searchable form integrated with the LDC Catalog so that searches for LRs reveal all related papers and vice-versa, (4) permitting paper authors to add their own citations to the database after validation and, finally, (5) further promoting the resource.

The LDC papers database, while still in-progress, has begun to address the motivations behind its creation. Just over half of LDC's LRs have been searched for and almost 8000 unique research papers have been located. These papers span two decades of research and have been presented at scores of academic conferences or published in various scholarly journals. The papers database has confirmed that those LRs that are licensed the most frequently by a research community are those that are also utilized most frequently for research.

Once the papers database has integrated into LDC's catalog, it will be utilized by LR users to locate all research related to an LR and by LR authors to help promote their LR and track usage. The facility with which researchers can exploit an LR and build upon prior work will increase, thereby lowering barriers to language-related education, research and technology development.

[1] http://www.ldc.upenn.edu/Catalog/topten.jsp

[2] LDC's Top Ten LRs were primarily published in the early 1990's.

8. References

Bird, Steven & Gary Simons (2000). A Survey of the State of the Art in Digital Language Documentation and Description, http://www.language-archives.org/documents/survey.html, Draft: 5

Brants, Thorsten & Alex Franz (2006). Web 1T 5-gram Version 1 (LDC2006T13). Linguistic Data Consortium, Philadelphia, PA.

Campbell, Joseph & Alan Higgins (1994). YOHO Speaker Verification (LDC94S16). Linguistic Data Consortium, Philadelphia, PA.

Charniak, Eugene et al. (2000). BLLIP 1987-89 WSJ Corpus Release 1 (LDC2000T43). Linguistic Data Consortium, Philadelphia, PA.

Chinchor, Nancy (2001). Message Understanding Conference (MUC) 7 (LDC2001T02). Linguistic Data Consortium, Philadelphia, PA.

Cieri, Christopher (2009). A Road Map for Interoperable Language Resource Metadata. SILT-FlareNet Workshop.Waltham, MA.

Garofolo, John S., et al. (1993a). TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1).
Linguistic Data Consortium, Philadelphia, PA.

Garofolo, John S., et al. (1993b). CSR-I (WSJ0) Complete (LDC93S6A). Linguistic Data Consortium, Philadelphia, PA.

Godfrey, John J. & Edward Holliman. (1997). Switchboard-1 Release 2 (LDC97S62). Linguistic Data Consortium, Philadelphia, PA.

Harman, Donna & Mark Liberman (1993). TIPSTER Complete (LDC93T3A). Linguistic Data Consortium, Philadelphia, PA.

Mitchell, Marcus, Beatrice Santorini, & Mary Ann Marcinkiewicz (1995). Treebank-2 (LDC95T7). Linguistic Data Consortium, Philadelphia, PA.

Mitchell, Marcus, et al. (1999). Treebank-3 (LDC99T42). Linguistic Data Consortium, Philadelphia, PA.

Robinson, Tony et al. (1995). WSJCAM0 Cambridge Read News (LDC95S24). Linguistic Data Consortium, Philadelphia, PA.