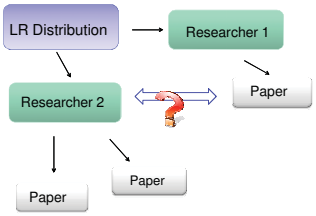
LDC Language Resource Papers: Building a Bibliographic Database

Eleftheria Ahtaridis, Christopher Cieri, Denise DiPersio
{ilya, ccieri, dipersio}@ldc.upenn.edu

Motivation

- LDC distributes over 500 LRs
- No single repository for all academic papers associated with an LR
- Little acknowledgment of LR efforts in papers
- No easy way to track LR impact



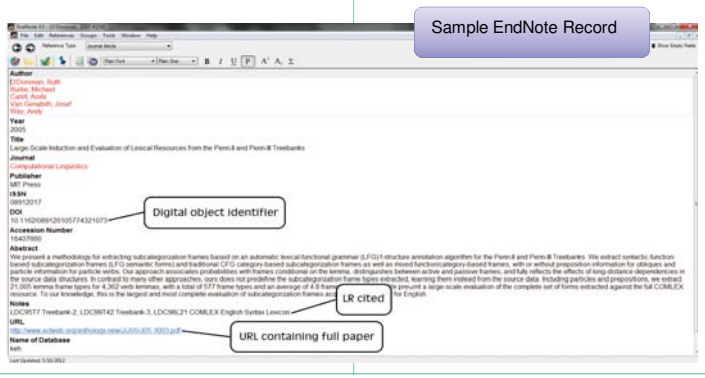
In 2009, LDC begun to collect metadata about papers that introduce, describe, or rely upon LRs in the LDC Catalog

Goal

- To create a database integrated with the LDC Catalog with bidirectional links between each LR and the papers mentioning them

Methodology

- LR Selection
  - Team identified major LR types – broadcast news speech, treebanks, news text etc – which were selected from across the entire LDC catalog
- Search Process
  - Conducted web searches for papers using Google Scholar® and CiteSeerX®
- Format
  - Papers archived using Endnote ® bibliographic software
  - At minimum each record includes: Author(s), Title, Year (of Publication or Conference), Journal Name or Conference Name, Abstract, URL, and LR(s) used



Problems Encountered

(1) Uncertainty about which specific LR was used
- Missing or inadequate citations
- Papers referred to LR by data source (ie. Wall Street Journal) and not official name

(2) LRs not cited in reference section of paper
- LR citations not standard in most research communities using LRs

(3) No mention of LDC either in the reference section or paper body
- LDC's suggested citation format includes LR title and author(s), year of publication, and lists LDC as publisher
- *Sample suggested citation:*
  R.H. Baayen, R. Piepenbrock and L. Gulikers 1996
  CELEX2
  Linguistic Data Consortium, Philadelphia

(4) Multiple URLs for papers found
- Google Scholar results often links to 'reader-pays' archives
- Avoided linking to personal web pages
- Attempted to locate freely available version
- When possible, chose URL of database like ACL Anthology, which provides free full text access

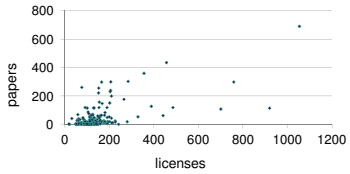
Progress to date

- Over 8000 references representing an in-depth search of about 55% of the LRs in LDC's catalog
  - Focused on published LRs
  - Papers describing only unpublished LRs were excluded from the database

Preliminary Analysis

- Almost 300 LRs, over half of LRs in LDC's Catalog, have been searched for extensively
  - Do most frequently licensed LRs appear most frequently in academic papers?

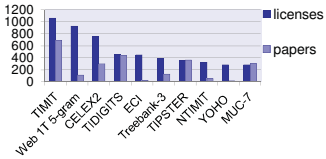
License Count Versus Paper Counts for LRs



- As expected, the more times an LR has been licensed, the more that LR is used - correlation coefficient .633

The number of papers for LDC's most licensed LRs, the "Top Ten" was examined
- In most cases, the more users of an LR, the more papers about that LR

License and Paper Counts for LDC's Top Ten LRs



- Of those LRs searched for extensively, average of 38 papers per LR located

Future Work

- Papers database resides on LDC's network. Remaining steps include:
  - (1) Completing initial search for all LDC LRs
  - (2) Extending OLAC or similar metadata repository to include links among LRs
  - (3) Converting EndNote database into web searchable form integrated with LDC catalog
  - (4) Permitting paper authors to add their own citations

Conclusion

LDC papers database, while still in progress, has begun to address the motivations behind its creation:
- LR users learn what research has been done which informs their future work
- LR creators gain feedback on their research
- Papers authors acknowledged for their work