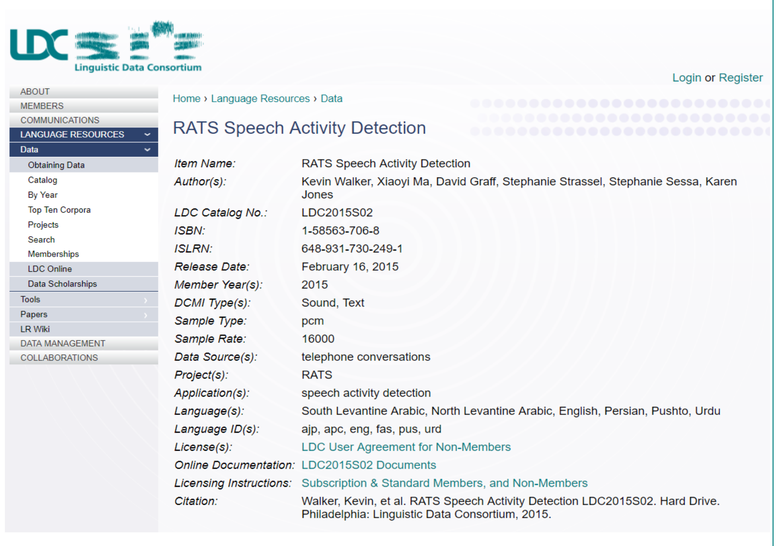
The LDC Catalog:
A Curated Repository of Language Resources

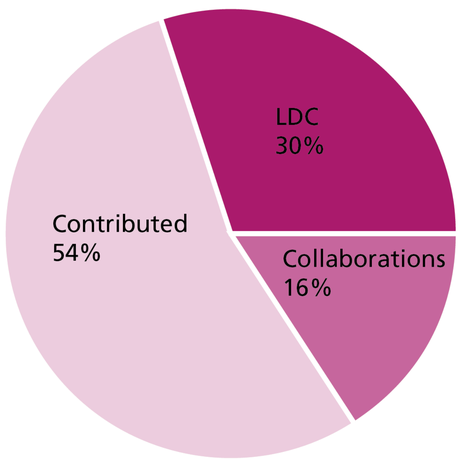
Denise DiPersio, Daniel Jaquette

Linguistic Data Consortium

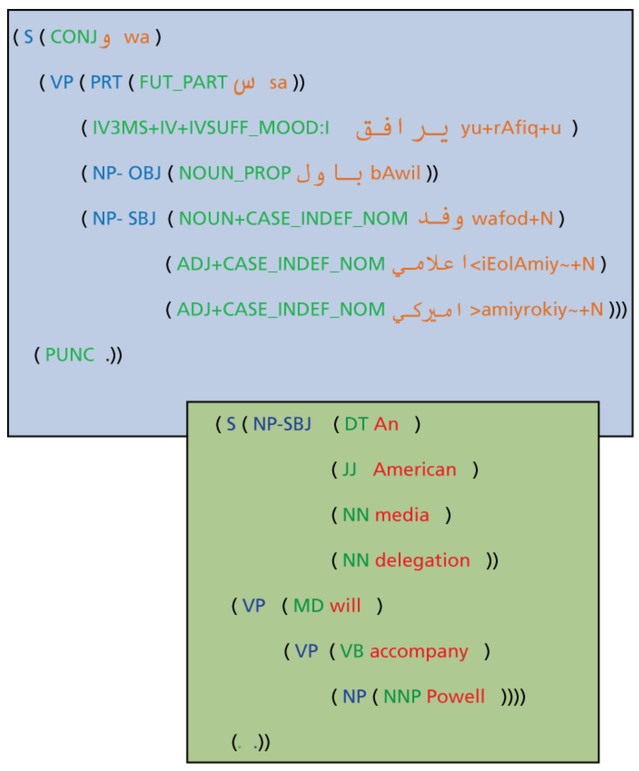
- The Linguistic Data Consortium (LDC) is an open, not-for-profit consortium of universities, libraries, and corporate and government research laboratories
  - Founded in 1992 as a permanent archive and distribution point for language resources (LRs)
  - Hosted at the University of Pennsylvania
  - A data center that creates, shares and archives corpora (data sets), software and specifications
- The LDC Catalog was one of the first public digital language resource repositories
  - Four basic LR types: lexicon, speech, text, video
  - More than 800 public holdings and grows by 36-40 releases annually
  - Includes corpora deposited under data management plans
  - LDC also develops and supplies LRs for sponsored projects and common task evaluations



- LDC's metadata schema is based off Dublin Core as modified for LRs by OLAC (Open Language Archives Community)
  - Discoverability: title, authors, persistent identifiers, citation
  - Descriptive: language, applications
  - Technical: data source, data type
  - Administrative: licensing, online documentation
- Multiple licensing options for compatibility with source data and data provider restrictions
  - Language-related research, education and technology development requires the use of "real world" data created for other purposes
    - LDC is a rights intermediary
    - LDC agreements with data providers (e.g., newswire, broadcast, web data) protect the community use case
    - Safeguards for providers' property rights, privacy, ethical concerns
- Regulatory compliance expertise
  - IRB for human subjects collections
  - Privacy
  - Export controls



- The LDC Catalog is a community resource
  - Donations and collaborative data sets comprise the majority of catalog resources
  - Global distribution network
  - Demonstrated research impact: 13,000+ identified research papers about LDC resources
  - Data by the numbers:
    - 175,000 copies
    - 6,000 unique organizations
    - 100 countries
- Most data distributed by LDC is annotated for a particular application
  - Data formats can affect downstream usability
  - Consortium guidelines for standard formats that work across platforms and consistent directory structures that promote easy reference
  - Parallel English-Arabic Treebank annotation sample (syntactic annotation to improve machine parsing of text)



Parallel Arabic and English Treebanked Sentence

- Majority of data is delivered via download directly from LDC
  - Physical media (disc, hard drive, flash drive) available depending on the needs of the data set or user
    - In house media hardware allows for automated large scale replication

- Data responsibility is at the core of LDC's mission to make language resources broadly available
- Comprehensive submission review
  - Quality reviews of formats, directory structure
  - Documentation reviews
  - Application of descriptive metadata
- Infrastructures and processes for reviewing, storing and distributing resources over the long-term
  - In-house storage (200TB+)
  - Redundant back-up systems
  - Automatic checksum, integrity checks over archive
  - Data migration to new formats, platforms, storage media following best practices
- All archived data (1992-present) accessible and discoverable
  - Catalog search functions, metadata mirrors in third-party repositories
  - New releases announced in LDC newsletter, website, social media
  - LDC user account management functions allow users to license data and join the Consortium online
- Procedures in place for data management plan development and execution
  - Assistance for budget development and distribution plan
  - Online submissions form
- A trustworthy repository
  - CoreTrustSeal awarded
- Submissions, curation and archive workflow at a glance: