More Data and Tools for More Languages and Research Areas:
A Progress Report on LDC Activities

Christopher Cieri, Mark Liberman
Linguistic Data Consortium

{ccieri|myl}@ldc.upenn.edu

Background

- **Changes in language resource landscape**
  - **Need for greater volume, growing number of languages, increasingly sophisticated annotation**
  - **System capacity grows independently giving partial match of supply, demand**
- **Advances in computing give**
  - **individual researchers ability to build corpora, address under-represented languages, disciplines.**
  - **data centers ability to store, process greater volumes of data at greater speeds**
- **Language resource production occasionally outpaces use**
  - **DARPA TIDES, EARS groups in machine translation and speech-to-text**
  - **focus shifed to source variation, annotation richness, quality, coordination**
- **As technologies approach human performance improving quality, understanding natural limits of human annotation performance become very important**
- **New communities, interdisciplinary teams adopt LR sharing, demand simple, adaptive access to data & flexible standards.**
- **Worldwide computing growth increases diversity of languages represented, demand for technologies and thus LR**

LDC Model

- LDC is Consortium, a group of organizations
  - 100-200 members join each year
  - 2019 organizations license data
  - management staff now 44 FT + 65 PT staff in Philadelphia
- Annual Membership
  - Online: access to subset of data included in LDC Online
  - Standard: up to 16 corpora per membership year, discounted licenses of data from previous years
  - Subscription (now 23%): 2 copies of each release shipped automatically, with priority (3 released/month)
- All tools, specifications and some data distributed without fee
- Members support consortium
  - through membership & license fees, trade, other special arrangements
- Much data available to non-members via licenses
- Benefits
  - broad data distribution with uniform licensing across research communities
  - funding agencies avoid distribution costs
  - users receive vast amount of data; avoid enormous development costs
    - » development costs of many corpora 1, 2 or even 3 orders of magnitude greater than NFP membership fee

LDC Activities

- **Resource Distribution (Cole)**
  - 31,300 (↑39%) copies of
  - 558 (↑94%) corpora + 3/month
  - 2019 (↑17%) organizations in 93 (↑4%) countries
  - Membership, licensing fees support this activity completely.
- **Intellectual Property Rights Management**
- **Data Collection**
  - news text
  - parallel text (Ma)
  - blogs (Strassel)
  - zines (Strassel)
  - newsgroups
  - broadcast news and broadcast talk (Strassel)
  - telephone conversation (Cieri)
  - meetings (Maeda)
  - read and prompted speech (Maeda)

LDC Activities

- **Annotation**
  - transcription (Strassel, Maamouri, Graff)
  - time-alignment
  - turn and word segmentation (Maeda)
  - morphological (Maamouri)
  - part-of-speech (Maamouri)
  - gloss (Maamouri)
  - syntactic (Bies, Maamouri)
  - semantic
  - discourse
  - disfluency
  - topic relevance
  - identification and classification of
    - » entities
    - » relations
    - » events
    - » co-reference
  - summarization
  - translation and multiple translation (Ma)
  - document, sentence, word level alignment of translation (Ma)

LDC Activities

- **Lexicon Building**
  - pronunciation (Graff), morphological (Maamouri), translation
- **Tools**
  - Transcriber,
  - MultiTrans & TableTrans
  - Buckwalter Arabic Morphological Analyzer
  - BITS: Bilingual Internet Text Search (Ma), Champollion (Ma)
  - XTrans: multichannel transcription (Maeda)
- **Infrastructure Building**
  - OLAC: Open Language Archives Community
  - Annotation Graph Toolkit (Maeda)
  - SPHERE Utilities
  - annotation workflow systems
- **Standards and Best Practices**
  - Topic Detection and Tracking v1.4, Entity Annotation Guidelines v2.5, Relation Annotation Guidelines v3.6, Simple MDE v6.2
- **Data Resource Coordination**
  - common task programs, outsourcing (ELRA MED Center, Arabic Transcription)
- **Consulting and Training**
- **Hosting and Maintaining research fora**
  - Talkbank Workshops, LDC Institute

Publications

- 2004 publication increases from 2 to 3/month to reduce 2 year backlog
- Dialogue Systems: 2000 & 2001 DARPA Communicator Dialogue Act
- Speaker Recognition: Switchboard Cellular II, 2002 NIST Speaker Recognition Evaluation
- Speech recognition:
  - Articulation Index (Jon Wright): multiple speakers producing <=2000 actual/nonsense English syllables
  - Transcribed broadcast news in Czech (JHU & U. W. Bohemia) and English
  - Transcribed telephone conversations in Mandarin (HKUST), English and Levantine Arabic
    - » Levantine requiring the invention of a writing system
  - Transcribed meeting audio in the ICSI, ISL (CMU) and NIST corpora
  - Read Speech: Speech Controlled Computing (Maeda), METU Turkish Speech, NATO Native, Non-Native
- DARPA EARS researched conversational disfluency/repair and identification of sentence-like units in speech to enhance automatic transcripts for downstream processing and improve readability
  - MDE RT-03 and RT-04 Training Speech and Annotations.
- Agreement with Center for Technology Enhanced Language Learning (CTELL) in U.S. Military Academy Department of Foreign Languages provided read speech corpora in Arabic, Russian, Croatian, English with additional publications planned
- Agreement (Jan van Santen, Mary Harper) with Oregon Health & Science University, Center for Spoken Language Understanding (CSLU) permits LDC to publish CSLU corpora for non-commercial education, research & technology development
  - CSLU Voices, CSLU: Spelled and Spoken Words, CSLU: Speaker Recognition V1.1, CSLU:Spoltech Brazilian Portuguese V1.0 remainder released over the coming months

Publications

- **Information Retrieval**
  - TDT4 English, Chinese, Arabic broadcast news & news text annotated for topic relevance
  - 2004 HARD text, topics and relevance annotations used in the TREC HARD track.
- **Information Extraction: ACE, DARPA TIDES provided English, Chinese, Arabic data annotated for entities, relations, events, co-reference**
  - 2003, 2004 and 2005 ACE/TIDES Multilingual Training
  - 2004 ACE Time Normalization English
  - BBN Pronoun Coreference and Entity Type, Timebank 1.2 (Pustejovsky, et. al.)
- **Machine Translation:**
  - 2 Arabic-English, 2 Chinese-English parallel news text
  - Hong Kong Parallel Text laws, press releases and Hansards,
  - 4 Chinese multiple translation corpora
  - Chinese-English name translation list
- **Morphological, Syntactic, Semantic analysis:**
  - Buckwalter Arabic Morphological Analyzer
  - Penn Chinese and Korean Treebanks (UPenn and Martha Palmer)
  - LDC Arabic Treebank, English-Arabic Treebank
  - Prague Arabic and Czech-English Dependency Treebanks (Charles U.)
  - Proposition Banks in English, Chinese Korean (UPenn and Martha Palmer)
- **TalkBank (UPenn, CMU, NSF) final 2 years focused on data creation**
  - Klex: Korean Finite-State Lexical Transducer, Morphologically Annotated Korean Text (NaRae Han, Mike Maxwell, CASL), Santa Barbara Corpora of Spoken American English III, IV, Field Recordings of Vervet Monkey Calls, FORM1 Kinematic Gestures (Craig Martell).
- **Other**
  - Mawukakan Lexicon (Moussa Bamba), Discourse Treebank (Florian Wolf), Arabic, Chinese and English Gigawords, 2eds., American National Corpus, 2nd release (Nancy Ide, Randi Reppen and ANCC

Projects

- DARPA TIDES concluded in 2005
  - Gigaword news text, annotations for topic relevance, entities, relation, events, coreference, parallel, translated, multiply translated text, summaries, English, Arabic, Chinese.
- DARPA EARS concluded in 2005
  - transcribed audio of conversational telephone speech (4000h English, 350h Mandarin, 260h Levantine Arabic, MDE annotation that identifies speakers, syntactic/semantic units and disfluencies/repairs
- DARPA GALE (Global Autonomous Language Exploitation)
  - transcribed broadcast, telephone conversations, news text, news groups, blogs translated, aligned at sentence and word levels, annotated for syntactic structure, propositional content, distilled into structured information.
- Mixer & Cross Channel (FBI, DoD, ITIC)
  - <=30 calls from 600 bilingual subjects, five languages, >=4 unique handsets and/or 9 different sensors
- Transcript Reading (FBI)
  - record each of 100 Mixer subjects reading samples of their own and others' transcripts via
- Language Variation and Dialect Identification
  - 100 telephone conversations in each of 32 linguistic varieties audited for language
- Automatic Content Extraction
  - English, Chinese, Arabic text from written, spoken sources annotated for entities, relations, events, co-reference.
- Less Commonly Taught Language (REFLEX)
  - resource kits for LCTLs including monolingual & parallel news text, bilingual lexicons, encoding converters, word & sentence segmenters, POS tagsets and taggers, morphological analyzers and tagged text, named-entity tagger and tagged text, personal name transliterator and grammatical sketch
- Spoken Language Treebanks for Levantine (Maamouri) and English (Bies)
- NIST Meeting Evaluation, TREC Video/VACE

Outreach via LDC Online

- LDC Data
  - Organized, formatted to accommodate HLT communities
  - Challenging for interactive use, annotation, non-technical researchers
- Solution LDC Online
  - Originally created in mid 90s with finds from NS; became brittle
  - Re-designed and re-built primarily local funding
- New approach provides access by language and data type not corpus
  - Brown Corpus, American English Spoken Lexicon are exceptions
- Indices
  - 500,000,000 words of Arabic
  - 1,400,000,000 Chinese characters
  - 2,600,000,000 words of English news text
  - 1500 hours (26,000,000 words) transcribed English conversations
  - news text indexed at word level, searchable metadata for date, source
  - conversations also tagged for project, topic, speakerID, sex, geographic region raised, age group, level of education
- Search Engine (Mike Schultz) optimized for speed and completeness
  - keyword and phrase search in text and in metadata fields
  - search terms combined with Boolean AND, OR and NOT
  - wildcards
  - relevancy ranking
  - search returns are keyword-in-context or, when possible, full text.
- Available to all LDC members
  - Plus 10,000,000 words, 10,000 documents available to registered non-members for non-commercial research at no cost

Outreach via LDC Online

Index	Documents	Words	Unique Words	Fields
Arabic News Text	1,692,835	489,729,594	3,188,548	▪ headline (type: indexed) – date (type: numeric) ▪ Start: 0 ▪ End: 20159109 – source (type: enum) ▪ afp (Agence France Presse) ▪ hyt (Al Hayat) ▪ nhr (An Nahar) ▪ umh (Ummah) ▪ xin (Xinhua News Agency)
Chinese News Text	3,087,084	1,368,064,442	844,985	▪ headline (type: indexed) – date (type: numeric) ▪ Start: 19910101 ▪ End: 20051231 – source (type: enum) ▪ afp (Agence France Presse) ▪ cna (Central News Agency) ▪ xin (Xinhua News Agency) ▪ zbn (Zaobao News)
English News Text	5,939,155	2,559,992,056	3,058,188	▪ headline (type: indexed) – date (type: numeric) ▪ Start: 19940512 ▪ End: 20051231 – source (type: enum) ▪ afp (Agence France Presse) ▪ apw (Associated Press) ▪ cna (Central News Agency) ▪ ltw (LA Times / Wash. Post) ▪ nyt (New York Times) ▪ sln (Salon.com) ▪ umh (Ummah) ▪ xin (Xinhua News Agency)

LDC Online: Search Results

English News Text | the | search

results view: ○ list ◉ tabular

Your search for **the** returned **151337745 hits** in **5733058 documents** . Refine your search now or start a **new query**.

#	Document ID	Headline
1	AFP_ENG_19940512.0003	Tributes pour in for late British
2	AFP_ENG_19940512.0004	France rules out participation ir
3	AFP_ENG_19940512.0005	Chinese dissidents in US favor
4	AFP_ENG_19940512.0006	Nagorno Karabakh hit by furth
5	AFP_ENG_19940512.0008	This restart looks good by Jim
6	AFP_ENG_19940512.0010	Aziz lobbies against oil embarg
7	AFP_ENG_19940512.0011	RAF planes airlift evacuees out
8	AFP_ENG_19940512.0012	South African shoots ahead in
9	AFP_ENG_19940512.0013	US Senate vote on Bonsia "inte
10	AFP_ENG_19940512.0014	(repetition) RAF planes airlift e
11	AFP_ENG_19940512.0016	Rightwing extremists chase Tu
12	AFP_ENG_19940512.0017	Spurs payments charge shock
13	AFP_ENG_19940512.0018	Radioactive capsule stolen in U
14	AFP_ENG_19940512.0020	ITT to open 750 million dollar c
15	AFP_ENG_19940512.0021	English cricket scores
16	AFP_ENG_19940512.0022	Zhirinovsky quizzes Russian PM
17	AFP_ENG_19940512.0023	World Cup organizers still fight
18	AFP_ENG_19940512.0026	India puts prize on Bombay bo
19	AFP_ENG_19940512.0027	Taiwan to allow massacre to fe
20	AFP_ENG_19940512.0028	China announces new minister

Keyword and Boolean Searching

Simple keyword searching is supported as well as standard boolean query syntax allowing for arbitrary combinations of AND's (&) OR's (|) and NOT's (!). Grouping is acomplished using parentheses.

- `cat`

Two items may be anded together using an implied AND or by an explicit use of &:

- `cat & dog`
- `cat dog`

Terms may be OR'ed together using the | operator. Documents containing either or both terms will be returned.

- `pants | trousers`

Parenthesis may be used to impose precidence on the operators. Otherwise left to right precidence is employed.

- `designer & (pants | trousers)`

The NOT (!) operator excludes documents from your search results but must be used in conjunction with another term. In other words, it must be used in an AND-NOT context.

- `cat &! dog` is equivalent to `!dog & cat`

Phrase Searching

Phrase searching is supported through the use of quotation marks. A hit occurs when all the terms of the phrase occur adjacent to each other and in the correct order. The use of a phrase causes the results to display these 'hits' instead of just the headlines of the documents containing them. Phrases can be used in boolean expressions.

- `"dog catcher"`
- `"dog catcher" &! cat`

Only the hits corresponding to the first phrase within a query are returned in the results lists.

- `"dog catcher" & "New York City"`
- `"New York City" & "dog catcher""`

The following two queries emphasize the difference between document based results and hit based results. Although the queries are identical, the number of entries returned differ. The number of hits is always greater than or equal to the number of documents since documents may have one or more hits.

- `Hosokawa & source:nyt` document mode

Plans

- **Maintain important role in language resource creation and distribution**
  - **support distribution efforts via memberships, data licenses**
  - **provide increasing support for local initiatives**
- **Extend outreach**
  - **new research communities**
  - **technology evaluations campaigns worldwide**
  - **commercial ventures requiring specialized corpora**
- **Efficiency**
  - **make better use of technologies that are based upon LDC data**
  - **simplify production through efficiency and outsourcing**
- **Increase research activities, collaboration**
- **Expand provision of tools, specifications, training to members**