# COCOSDA: North American Report

Christopher Cieri
ccieri@ldc.upenn.edu
University of Pennsylvania
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia PA. 19104, USA

# NIST Evaluations

| | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 | 2002 | 2001 | 2000 | 1999 | 1998 | 1997 | 1996 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GALE Translation | ■ | ■ | ■ | | | | | | | | | | |
| Rich Transcription | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | |
| Spoken Term Detection | | | ■ | | | | | | | | | | |
| Speaker Recognition | ■ | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Language Recognition | | ■ | | ■ | | ■ | | | | | | | ■ |
| Topic Detection and Tracking | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | |
| Information Extraction | | | | | | | ■ | ■ | ■ | ■ | | | |
| Communicator | | | | | | | | ■ | ■ | ■ | | | |
| Conversational Telephone Recognition | | | | | | | | ■ | ■ | | ■ | ■ | |
| Spoken Document Retrieval | | | | | | | | | ■ | ■ | ■ | ■ | |
| Broadcast New Recognition | | | | | | | | | | ■ | ■ | ■ | ■ |

# Sample Projects

- DARPA GALE (Global Autonomous Language Exploitation)
  - supports multilingual transcription, translation into English and distillation of text into structured information
  - text (news, newsgroup, blog), transcribed speech (broadcst news and conversation) translated and aligned at sentence and sub-sentence level, annotations for syntactic structure & propositional content, distillation into structured information.
  - English, Mandarin and Arabic
- MADCAT
  - supports systems that perform OCR (,LR) and MT of handwritten, printed and hybrid text
  - varying scribe, text type, writing instrument, time, speed of writing, paper quality
  - first language Arabic
- Mixer Phases 1-5
  - support robust speaker recognition technologies
  - multigenre: conversational telephone speech, transcript reading, face-to-face interviews
  - multilingual: Arabic, English, Mandarin, Russian, Spanish
  - multichannel: lavalier on the subject and interviewer, Etymotic Link-It micro-array, podium, PZM, studio, hanging conference room, camcorder, 4 studio mics at varying distances from subject, microphone array, head mounted mic used only for brief telephone calls
- LVDID (Language Variation and Dialect Identification)
  - >100 conversations in each of a dozen linguistic varieties
  - ongoing collection in another 20 varieties with all calls audited for sound quality and language * Chinese: Shanghai Wu Dialect (Shanghai Hua)
  - Chinese: Southern Min Dialect/Taiwanese (Min Nan Hua/Taiwan Hua), Chinese: Cantonese (Guangdong Hua/Yue Yu), Georgian, Hindi, Farsi, Tamil, Japanese, Korean, Italian, Punjabi, Aceh, Amharic, Bengali, Burmese, Chechen, Guarani, Khmer, Lao, Tagalog, Thai, Tigrigna, Urdu, Uzbek
- REFLEX-LCTL (Less Commonly Taught Languages) [Simpson, et. al.]
  - supports multiple technologies for LCTLS especially extraction and translation
  - 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
  - Amazigh (Berber), Bengali, Hungarian, Pashto, Punjabi, Kurdish, Tagalog, Tamil, Thai, Tigrigna, Urdu, Uzbek and Yoruba

# Projects

- TRANSTAC – STS translation, limited domain, portable platform, Arabic, Persian
- HAVIC - web video collected, classified and annotated
- TREC Video - broadcast video, key frames, transcripts
- Mixer Greybeard - multiple telephone conversations from subjects in previous studies
- OLAC - ongoing development of the Open Language Archives Community

# JHU Summer Workshop

- 2004
  - Joint Visual-Text Modeling
  - Landmark Based Speech Recognition
  - Dialectal Chinese Speech Recognition
- 2005
  - Parsing Arabic Dialects
  - Parsing and Spoken Structural Event Detection
  - Statistical Machine Translation by Parsing
- 2006
  - Articulatory Feature-based Speech Recognition
  - Open Source Toolkit for Statistical Machine Translation
- 2007
  - Exploiting Lexical & Encyclopedic Resources For Entity Disambiguation
  - Recovery from Model Inconsistency in Multilingual Speech Recognition
- 2008
  - Multilingual STD Finding and Testing New Pronunciations
  - Robust Speaker Recognition
  - Vocal Agin Explained by Vovcal Tract Modeling

# Other

- JHU Center for Excellence in Speech and Text processing
  - Gary Strong, Jim Baker
  - US Gov’t Funding planned through 2015
  - revolutionary advances in speech and text

# Language Resource Landscape

- Some speech technologies approaching human performance
  - quality
  - understanding natural limits of human performance

become very important

# LDC

- Linguistic Data Consortium established 1992
  - centralized location to distribute and archive language data
  - normalize and manage intellectual property rights and distribution practice
- Organized as a consortium, group of organizations, hosted by U. Penn.
- Management staff in Philadelphia: 45 FT & <= 65 PT employees
- Funding
  - DARPA seed funding covered operations + corpus creation
  - early support from NSF, NIST
  - required to be self-sufficient within 5 years (operation costs<= fees)
  - annual membership fees, data licenses
  - grant funding for specific resource creation, not maintenance
- Data comes from donations, funded projects at LDC or elsewhere, community initiatives and LDC initiatives.
- Expansion
  - 1995: collection, transcription activities, 1998: annotation,1999: tools and standards, 2002: coordinating multi-site efforts, sharing experience through publications, training
- LDC's mission as currently defined is to support language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards.
- Activities
  - resource distribution, intellectual property rights management, resource production
  - data collection, annotation, lexicon building
  - tool creation, infrastructure building
  - creation of best practices, consulting and training
  - corpus creation research, resource coordination

# LDC

- data collection
  - news text for LM
  - broadcast news and talk
  - telephone conversation, meetings, interviews
  - read and prompted speech
- annotation
  - quick and careful transcription
  - time-alignment and segmentation at the turn, sentence and word level
  - tagging of morphology, part-of-speech, gloss
  - syntactic annotation
  - discourse function and disfluency
  - categorization according to topic relevance
  - identification and classification of entities, relations, events and their co-reference
  - summarization of various lengths from 200 words down to titles
  - translation, multiple translation, translation quality control
  - alignment of translated text at the document, sentence and word levels
- lexicon building
  - pronunciation, morphological, translation

# Publications/Membership

- New Membership types:
  - Online: access to subset of data included in LDC Online
  - Standard: LDC Online plus may request licenses <= 16 corpora, discounted licenses of data from previous years, discounted extra copies of licensed data
  - Subscription: Standard Members but automatically receive 2 copies of all corpora on media as they are released
- Subscription memberships, added in 2005, now account for 23% of all members.
- Reorganized fee structure
- LDC currently adds 2-3 corpora to Catalog/month.
  - Membership and licensing fees support this activity completely
- LDC has distributed 53,580 (↑71%) copies of nearly 800 (↑43%) corpora and otherwise shared data with 2540 (↑26%) organizations in 67 countries.

# Publications

- Since last report, LDC added
  - 68 titles to Catalog + dozens of corpora for evaluation programs
- A sampling of those corpora includes:
  - Gigaword (billion word) News Text corpora in Arabic, Chinese, English, French, Spanish
  - broadcast news in Arabic, Korean
  - many contributions from Center for Spoken Language Understanding (CSLU)
    - Foreign Accented English, Apple Words and Phrases, Yes/No, Spelled and Spoken Words, Stories, Multilanguage Telephone Speech, Portland and National Cellular Telephone Speech, Names Release, Speaker Recognition, Spoltech Brazilian Portuguese and Voices
  - parallel text including Arabic Blogs (DARPA GALE)
  - STC-TIMIT: TIMIT data process through telephone network contributed by (Morales)
  - Urdu speech from the Army Research Labs
  - Speech in Korean and Spanish contributed by West Point
  - Conversational Telephone Speech in Levantine, Iraqi and Gulf Arabic
  - Broadcast News Parallel Text (LDC, MITRE)
  - Video, key frames and transcripts created by the TRECVID program
  - Broadband Prompted Speech in English and Turkish (Middle East Technical University)
  - Telephone Band Speech in Russian
  - Evaluation data from the NIST 2003 and 2004 Rich Transcription campaigns