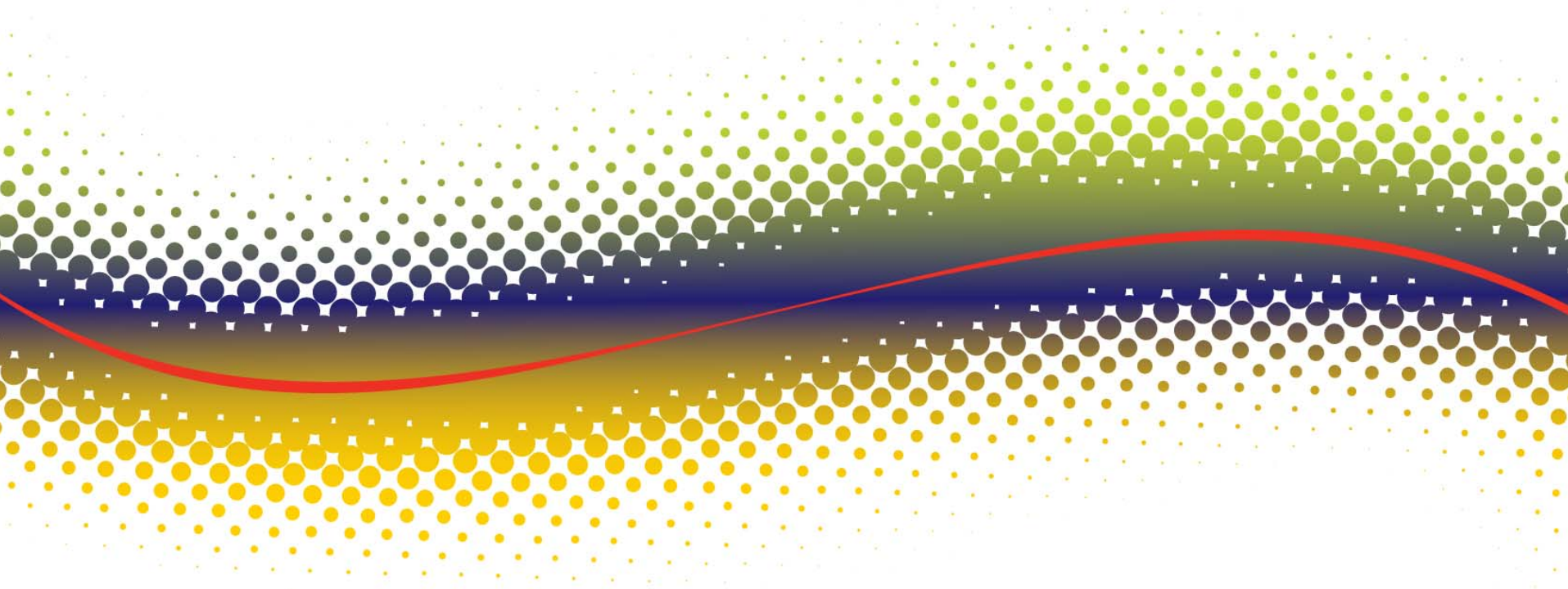
Linguistic Data Consortium

COCOSDA: Report for the Americas

Christopher Cieri

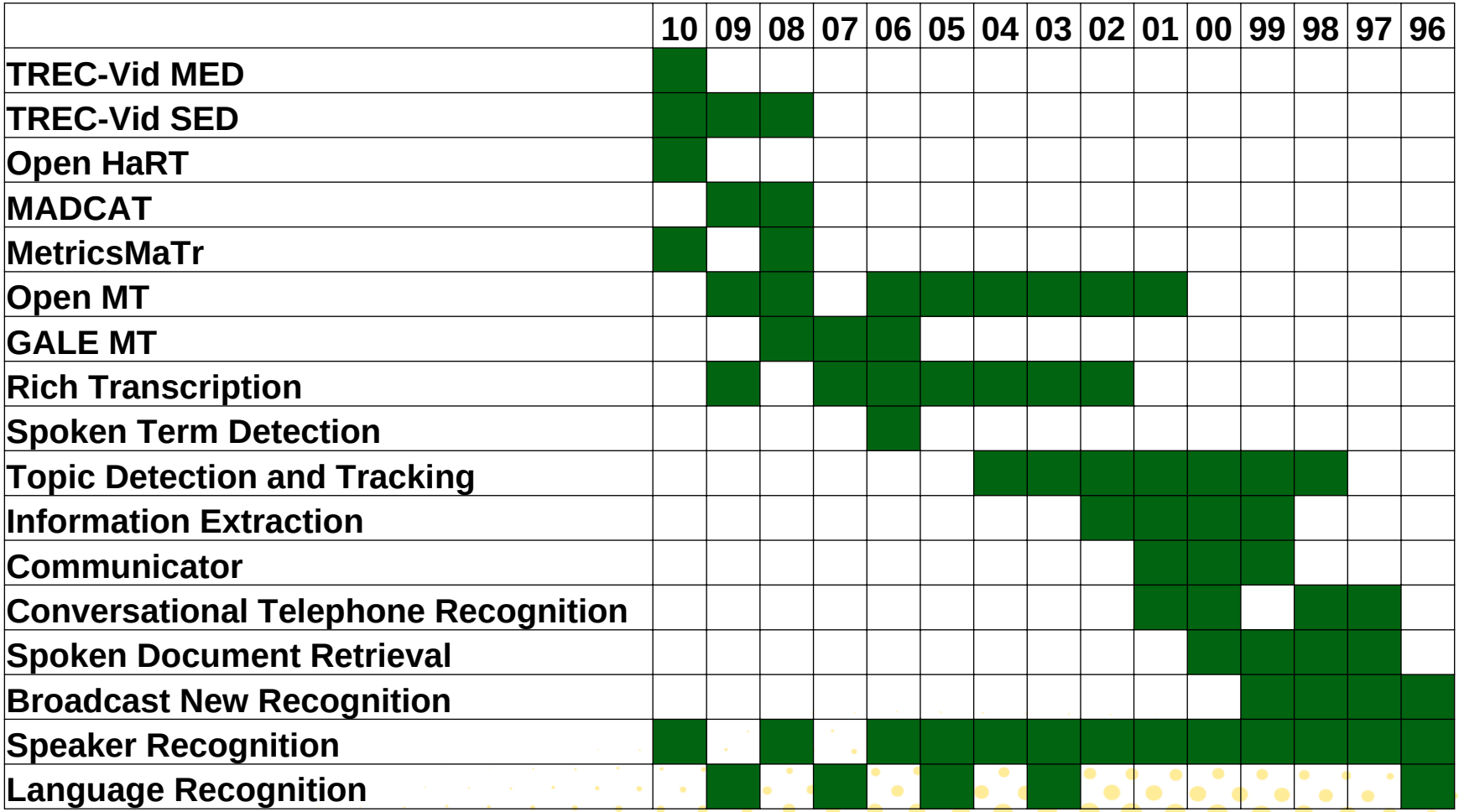
University of Pennsylvania, Linguistic Data Consortium

ccieri AT ldc.upenn.edu



Linguistic Data Consortium

NIST Evaluations



DARPA IPTO Language Processing

- GALE (Global Autonomous Language Exploitation)
  - multilingual transcription, translation into English, distillation of text into structured information
  - text (news, newsgroup, blog), transcribed speech (broadcast 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 and MT of handwritten, printed and hybrid text
  - varying scribe, text type, writing instrument, time, speed of writing, paper quality
  - first language Arabic
- TRANSTAC – STS translation, limited domain, portable platform, Arabic, Persian

DARPA IPTO Language Processing

- RATS (Robust Automatic Transcription of Speech)
  - Algorithmic Development and Signal Processing
    - Speech Activity Detection (SAD): speech, background noise, music
    - Language Identification (LID): language spoken in signal determined speech
    - Speaker Identification (SID): whether speaker is one of list of target speakers
    - Key Word Spotting (KWS): identify specific words or phrases from a list
  - Data Collection
    - data for SAD, LID, SID, KWS
    - must include data matrix
  - Evaluation

Data Matrix

		Language	Genre	Unit	Minimum Targeted Volume	Volume Released to Date	R1		R2		R3	Notes
							10.1.2007		4.4.2008		TBD	
Training	Collection	Arabic	NW	Words	*See Notes*	*See Notes*						All collected NW to be released in Gigaword 4 (2009) and/or ad hoc as required for specific tasks
			BN	Hours	1000	1103.561	372.797	*LDC2007E99*	730.764	*LDC2008E38*		Includes both LDC- and web-harvested audio
			BC	Hours	1000	950.616	434.771	*LDC2007E99*	515.845	*LDC2008E38*		Includes both LDC- and web-harvested audio
			WL	Words	10,000,000	98,729,851	26,296,828	*LDC2007E102*	72,433,023	*LDC2008E41*		
			NG	Words	10,000,000	70,784,544	27,838,327	*LDC2007E102*	42,946,217	*LDC2008E41*		
		Chinese	NW	Chars	*See Notes*	*See Notes*						All collected NW to be released in Gigaword 4 (2009) and/or ad hoc as required for specific tasks
			BN	Hours	1000	620.824	377.331	*LDC2007E99*	243.492	*LDC2008E38*		Includes both LDC- and web-harvested audio
			BC	Hours	1000	697.515	304.718	*LDC2007E99*	392.797	*LDC2008E38*		Includes both LDC- and web-harvested audio
			WL	Chars	15,000,000	206,361,199	185,550,342	*LDC2007E102*	20,810,857	*LDC2008E41*		
			NG	Chars	15,000,000	271,207,196	164,658,186	*LDC2007E102*	106,549,010	*LDC2008E41*		
		English	NW	Words	*See Notes*	*See Notes*						All collected NW to be released in Gigaword 4 (2009) and/or ad hoc as required for specific tasks
			BN	Hours	250	120.59	60.216	*LDC2007E99*	60.374	*LDC2008E38*		Includes both LDC- and web-harvested audio
			BC	Hours	250	120.637	60.234	*LDC2007E99*	60.403	*LDC2008E38*		Includes both LDC- and web-harvested audio
			WL	Words	10,000,000	4,515,219	3,024,785	*LDC2007E102*	1,490,434	*LDC2008E41*		
			NG	Words	10,000,000	827,115,809	400,201,443	*LDC2007E102*	426,914,366	*LDC2008E41*		
		Arabic	BN	Hours	500	449.699	305.593	*LDC2007E100*	144.106	*LDC2008E39*		Includes manual and web-harvested transcripts
			BC	Hours	500	543.49	380.938	*LDC2007E100*	162.552	*LDC2008E39*		Includes manual and web-harvested transcripts
			BN	Hours	500	341.138	293.252	*LDC2007E100*	47.886	*LDC2008E39*		Includes manual and web-harvested

Specifications

GALE: Task Specifications and Annotation Guidelines

Task specifications state needs and assumptions for each task, describe the process for collecting and/or selecting data for that task, define annotation and quality control procedures associated with the task, and describe the distribution formats for the resulting data. LDC's GALE tasks include

- **Collection**
  - **Broadcast data (news and talk shows)** (updated 4/1/2008)
    - **Broadcast auditing** (updated 4/1/2008)
  - **Web data (blogs and newsgroups)**
- **Transcription**
  - **XTrans (speech annotation toolkit) manual V3** (updated 10/11/2007)
- **Translation**
- **Word Alignment**
  - **Arabic Word Alignment V4.0** (updated 04/08/2009)
  - **Chinese Word Alignment V4.0** (updated 4/16/09)
  - **Chinese Tagging Guidelines V1.0** (updated 4/10/09)
- **XBanks**
  - **Arabic Treebank** (updated 1/15/2009)
  - **English-Arabic Treebank** (updated 4/9/2009)
- **Distillation**
  - **Phase 3 Training Data Annotation Guidelines V1.0** (updated 06/18/2008)
- **Evaluation Resources**
  - **Data Selection Guidelines V2.2** (updated 01/02/2007)
  - **Machine Translation Post Editing**
    - **Post Editing Guidelines V3.0.2** (updated 05/25/2007)
- **Resource Distribution**
  - **Resource Distribution** (updated 4/1/2008)

Linguistic Data Consortium

IARPA

- IARPA – "*Intelligence Advanced Research Projects Activity invests in high-risk/high-payoff research … cross-agency challenges … expertise from across the community …work transition strategies and plans …*"
- Smart Collection
  - BEST (Biometrics Exploitation Science & Technology)
    - multiple biometrics: face, ocular, voice
    - challenging collection conditions
- Incisive Analysis
  - ALADDIN (Automated Low-Level Analysis and Description of Diverse Intelligence Video)
  - SCIL (Socio-cultural Content in Language) – "*discovery of the social goals of members of a group by correlating these goals with the language they use*"

- Computing Research Infrastructure
  - Manually Annotated Sub-Corpus
    - MASC I: 82,000 words from O-ANC
    - manual annotations
      - WordNet senses
      - fulltext FrameNet frame annotations
      - validated annotations for token and sentence boundaries
      - part of speech
      - noun chunks, verb chunks
      - named entities
      - Penn Treebank syntactic annotations
      - includes texts from the Language Understanding Corpus
      - about half of the corpus was annotated in the ULA project
        - annotations for opinion, PropBank, and TimeML included in MASC I or forthcoming
      - annotations in LAF/GrAF
- Cyberinfrastructure

Department of Education

- IRS
  - Reading Assistance and Assessment Tools for Morphologically Complex Languages
    - Arabic, Nahuatl
    - browser, morphological analyzer, lexicon
    - modifications to accommodate terminology of classroom
    - curriculum analysis
  - Digital Dictionaries of Arabic Colloquial Varieties
    - GUP Dictionaries
    - Iraqi, Syrian, Moroccan
    - normalized Person-Arabic orthography, IPA pronunciations
    - new entries based on modern corpora
  - Survey of DOE funded dictionary projects
    - Doug Cooper

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 Aging Explained by Vocal Tract Modeling
- 2009
  - Parsing the Web: Large-Scale Syntactic Processing
  - Low Development Cost, High Quality Speech Recognition for New Languages and Domains
  - Unsupervised Acquisition of Lexical Knowledge from N-Grams

Miscellaneous

- Mixer Phases 6-7
  - support robust speaker recognition technologies
  - multigenre: conversational telephone speech, transcript reading, face-to-face interviews, repeating questions
  - multilingual: Arabic, English, Mandarin, Russian, Spanish
  - multichannel: lavalier on the subject and interviewer, podium, PZM, studio, hanging conference room, camcorder, studio mics at varying distances from subject, microphone array, head mounted mic used only for brief telephone calls
  - HVE, LVE
- LVDID (Language Variation and Dialect Identification)
  - >100 conversations in each of a dozen linguistic varieties
  - BNBS
- Mixer Greybeard - multiple telephone conversations from subjects in previous studies
- HAVIC - web video collected, classified and annotated
- TREC Video - broadcast video, key frames, transcripts

“South America”

- Fala Brasil - Reconhecimento de Voz para o Português Brasileiro from Federal University of Para, Brasil
  - LaPSAM v1.3 – acoustic model via HTK
  - LaPSLM v1.0 – n-gram language model via SRILM
  - UFPAdic.3.0 – 38 phone dictionary, based on SAMPA
  - LapsBenchMark1.4 – test corpus 700 sentences (35 speakers * 20 phrase each (25 male, 10 female)
  - TextCorpora1.5 – sentences used to train LM
  - LapsNews1.0 – 120K sentence web text corpus

“South America”

- The Spoltech Brazilian Portuguese (LDC2006S16)
  - prompted sentences and answers to questions
  - multiple regions of Brazil
  - 8080 utterances from 477
    - 2572 orthographic transcriptions
    - 5507 time-aligned phoneme-level transcriptions
- West Point Brazilian Portuguese Speech (LDC2008S04)
  - digital recordings of spoken Brazilian Portuguese collected by … CTELL to develop acoustic models for speech recognition systems … used to provide speech-recognition enhanced language learning courseware to government linguists and students”
  - read speech from 60 female and 68 male native and non-native speakers.
  - prompt script containing 296 sentences and phrases typically used in language learning situations.