Christopher Mark Cieri

Coordinates

Professional
ccieri@ldc.upenn.edu
http://www.ldc.upenn.edu/~ccieri
http://orcid.org/0000-0002-8509-6413

Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104
+1-215-573-5489

Education

Ph.D. University of Pennsylvania, Philadelphia, PA. May 2005
Doctor of Philosophy, in Linguistics. Dissertation: Modeling Phonological Variation in Multidialectal Italy. Areas of concentration: Sociolinguistics, Phonology, Phonetics, Historical and Educational Linguistics.

M.A. University of Pennsylvania, Philadelphia, PA. January 1986
Master of Arts in Linguistics. Concentration in Language Contact, Sociolinguistics, Phonology and Morphology. Thesis: Italian Lexical Items in the English Speech of Italo-Americans. GPA 3.88. Completed M.A. concurrently with B.A.

B.A. University of Pennsylvania, Philadelphia, PA. August 1985
Bachelor of Arts, Cum Laude and with Distinction in Linguistics. Other course work in Cognitive Psychology, Cultural Anthropology and Computer Science. GPA - general 3.40, major 3.83. Philadelphia Mayor's Scholar.

St. John Neumann High School, Philadelphia, PA. June 1981
National Merit Scholar Commendation, PHEAA Commendation, Philadelphia Distinguished Scholar, Dean's List, Faculty Distinction in Academics, Faculty Distinction in Activities, First Honors all semesters.

Research Interests

Linguistics: Quantitative analysis and modeling of linguistic variation, phonetics, phonology, morphology, dialectology, Italian and Italian dialects, language contact, corpus creation and linguistic annotation

Speech and Language Technology: Linguistic databases, Standards and tools for linguistic annotation, Computer-assisted linguistic analysis, Computer-assisted language learning.

Selected Research Grants

Principal Investigator or Co-Principal Investigator

- Linguistic Resources for Multilingual, Genre-Independent Language Technologies, BOLT, Department of Defense , HR0011-11-C-0145, September 2011 – May 2015, $6,850,490
- Vipertoxin, MPO, H98230-10-D-0041, September 2012 – March 2015, $5,247,718
- SI2-SSI: Web Services for Natural Language Applications, NSF via Brandies University, 3-02069, August 2102 – July 2015. $219,170

- ReORIENT: Resources for Operationally Relevant Information Extraction from Non-explicit Text, DEFT, DARPA, FA8750-13-2-0045, November 2012 –May 2017, $5,983,769
- Linguistic Resources for NIST Evaluations - TRECVid MED/MER, SB1341-13-CQ-0021, Department of Commerce, October 2013 – September 2014, $1,018,230
- Data Resources Aladdin Video Evaluations, IARPA, 2104-14042400005, May 2014 – October 2015, $749,928
- Linguistic Resources for Robust Automatic Transcription of Speech, DARPA RATS, August 2010 – February 2012, $3,271,926.
- Voice Data Collection, Linguistic Annotation and Distribution, Department of Defense, H98230-10-R-0293, June 2010- May 2013, $2,047,717.
- Mining a Year of Speech, NSF, Digging into Data Program, August 2010-July 2011, $98,899.
- Multilingual Automatic Document Classification Analysis and Translation, DARPA MADCAT, HR0011-08-1-004, December 2007–September 2013, $6,041,455.
- GALE: Integrated Linguistic Resources for Language Exploitation Technologies, DARPA, HR0011-03-1-0003, October 2005–June 2011, $27,987,854.
- Phanotics, MIT-LL, 7000013887, August 2007–February 2009, $118,000.
- Machine Reading, SAIC, July 2009-June 2014, $1,433,339.
- IDIQ: Indefinite-Delivery/Indefinite Quantity Provision of Linguistic Resources, NIST SB134107CQ0024, 9/1/2007–8/1/2012, $9,701,612.
- Teaching & Learning Aids for Linguistically Complex Languages, Department of Education IRSG, P017A0500401-01, 9/1/2005–12/31/2008, $467,000.
- Urdu Machine Translation, Department of the Army, W911NF-09-1-0422, September 2008—March 2009, $6,518.
- Seedling Corpus: Self-Sustaining RC ECP, PO#06-C-0025-02A, September 2008—December 2008.
- IRS 2008: Dialectal Arabic Dictionaries, Department of Education, P017A080044-1, July 2008—July 2009, $165,000.
- GeoParsing: Support for Pilot Evaluation of GeoParsing Technology, Insightful Corporation, RE2310, June 2006—December 2007, $45,907.
- LCTL: Less Commonly Taught Languages, Department of the Interior, NBCHC050085 & NBCHC050085, June 2005–June 2006, $1,409,932.
- ISLE/EMELD: National Science Foundation via Wayne State University, WSU#-1171-A1, 7/19/2001–6/30/2006, $493,548.
- LVDID: Language, Variety and Dialect Identification, U.S. Department of Defense, H98230-04-C-0511, September 2004–September 2008, $2,423,848.
- Reflex/MT: Machine Translation Corpora Creation, Department of the Interior, NBCH050083-0001 and NBCH050083-0001, 9/30/2005–9/29/2006, $216,460.
- Arabic Propbank, University of Colorado, 1542294, 10/1/2005–8/31/2006, $75,818.
- MMSR: Multilingual, Multi-channel Speaker Recognition, Massachusetts Institute of Technology, 0-201-06-62-2003 and 0-201-03-02-8003, July 2003–September 2005, $853,937.
- JHU Summer Workshop on Speech Parsing, Johns Hopkins University, 8506-03014, March 2005–September 2005, $65,361.
- Transcription & Distribution of RT-05 Meeting Speech Evaluation Data, Department of Commerce, SB134105W0651, June 2005–December 2005, $71,784.
- VACE: Video Analysis and Content Extraction, Department of the Interior, NBCHC040070-003, July 2005–June 2006, $198,088.
- TRECVid: Text Retrieval Conference – Video, Department of the Interior, NBCHC040034-0004, July 2005–June 2006, $198,675.
- Linguistic Resources for Automatic Content Extraction, Department of the Interior, NBCHC040172-0001, September 2004–September 2006, $506,557.

- Meeting Transcription, National Institute of Standards and Technology, NA1341-03-W-1210, September 2003–June 2004, $55,000.
- Foreign Language Newswire, Department of Defense, MDA904-03-C-0596, September 2003–April 2005, $122,462.
- Speech Controlled Computing, Microchip Corporation, October 2004–December 2004, $37,436.
- Networking Data Centers, National Science Foundation, IIS-9982201-005, September 2000–August 2006, $735,378.
- Talkbank, National Science Foundation KDI/SBE, BCS-998009, October 1999–September 30, 2005, $1,401,215.
- Digitized Broadcast News Distribution, Department of Defense, MDA904-03-C-0469, July 2003–July 2005, $141,522.
- Advanced Linguistic Resources and Infrastructure Supporting Human Language Technology Development (TIDES & EARS), DARPA, N66001-02-1-8904, $13,276,394.
- TIDES: Translingual Information Detection Extraction and Summarization, DARPA, N66001-00-2-8918, February 2000–February 2003, $2,792,698.
- Data to Support Speech-Processor Research & Development, Department of Defense, MDA904-00-C-2079, May 2000–September 2001, $472,619.
- Tactical Speech Corpus, Massachusetts Institute of Technology, 536144, 7/1/2000-9/30/2000, $16,579
- Improved Speech and Text Data Resources, National Science Foundation, IRI-95-28687-05, January 1999–March 2001, $3,894,069.
- Sun Academic Equipment Grant, Sun Microsystems, 7826-990 237-US, 1999, $90,000.
- Oleada, Department of Defense, MDA904-97-C-0307, March 1997–December 2000, $735,238.
- LDC Fund (income from consortium members and data sales): 1998-present, $20,184,479.

Research Specialist

- Multilingual, Multimedia Training Software (U.S. Air Force Academy), Research Specialist, 4/1989-9/1989, Evaluated computer platforms (NeXT and PC with DVI) for development of multilingual, multimedia training software.
- Data Discrimination and Categorization (DOD), Research Specialist, 6/1988-9/1988, Developed prototype system to analyze unknown files and identify file type, language and topic.
- French Reading Tools (University of Pennsylvania Internal Grant), Research Assistant & contributor to proposal, 5/1987-8/1987, Developed software and lexical databases to aid students reading French including Old French.
- Phonological Toolbox/TextWorks (DOD MDA90485H0101), Research Specialist, 9/1983-12/1989, Managed development of the toolbox to help linguists analyze sound systems working from phonetic transcriptions; Managed creation of 4 electronic bilingual dictionaries and a 15 million word text base spanning 35 critical languages.
- Networks in Language Learning (National Foreign Language Center), Research Assistant, 2/1987-8/1987, Investigated local area networks in computer aided language learning; developed reading aid that also captured data on individual reading strategies.

Teaching Experience

Best Practices in Sociophonetic Field Recording: Digital Tools and Transcription. Lecture presented at Malcah Yaeger-Dror and Marianna DiPaolo, Workshop on Sociophonetics, LSA Summer Institute, Boulder, CO, July 2011.

Resources and Evaluation of Text and Speech Systems – 10th ELSNET Summer School on Language and Speech Communication, Odense, July 2002

Introduction to Language II - phonetics, phonology, morphology, historical linguistics, sociolinguistics, writing systems – University of Pennsylvania, College of General Studies: Fall 1996, Summer 1996, Fall 1987

Advanced English: The Grammar of English Poetry (two day seminar), The Pronunciation of American English (two day seminar) – Università dell'Aquila, Istituto Inglese, Spring 1988

Professional Experience

Executive Director, Linguistic Data Consortium: January 1998-present.
Director, Law School Computer Services: January 1990-December 1997.
Research Specialist, Language Analysis Center: September 1983-December 1989

Selected Professional Activities

Proposal Review Panels

- National Science Foundation, Linguistics, Human-Computer Interaction and Human Language and Communication programs: 2001, 2003, 2007-2010.
- Netherlands Organization for Scientific Research (NWO), 2012.
- Flemish Ministry of Science and Innovation via the Hercules Foundation, 2012

Journals

- LRE: Language Resources and Evaluation Journal: Advisory Board & Reviewer, 2005-present
- Speech Communication: Reviewer, 2014
- Penn Review of Linguistics: Reviewer

Conferences

- NAACL-HLT: Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, 2015, Reviewer
- COLING14: 14th Computational Linguistics Conference: Workshop on Open Infrastructures and Analysis Frameworks for HLT, 2014, Workshop Chair
- LREC: Language Resources and Evaluation Conference: Scientific Review Committee, biennial 2002-present. Workshops on: Language Resource Management and Sustainability, Collaborative Resource Development and Delivery, Indian Language Data Resources and Evaluation.
- CLC7: 7th Celtic Linguistics Conference, 2012, Scientific Committee.
- The LAW: Linguistic Annotation Workshop: Program Committee, 2007-2009, 2012
- LTC: Language & Technology Conference, 2011: Human Language Technologies as a Challenge for Computer Science and Linguistics
- IJCNLP - International Joint Conference on Natural Language Processing: Workshop on Language Resources, Technology and Services in the Sharing Paradigm, Scientific Committee, 2011
- ACL Program Committee, 2010
- EMNLP: Conference on Empirical Methods in Natural Language Processing, Reviewer and Best Reviewer Award Recipient, 2010
- EACL: European Chapter of the Association for Computational Linguistics, African Language Technologies Workshop: Reviewer, 2009.
- MEDAR: Mediterranean Arabic Language and Speech Technology Program, 2nd International Conference on Arabic Language Resources and Tools: Conference Program Committee and Scientific Committee, 2008
- ICGL: International Conference on Global Interoperability for Language Resources: Program Committee, 2007, 2008

- EMELD: Electronic Metastructure for Endangered Languages Data project, Digital Tools Summit in Linguistics: Organizing Committee, 2006, 2007
- IEEE Transactions on Audio, Speech and Language Processing: Reviewer, 2007
- NWAVE: New Ways of Analyzing Variation Conference: Editorial Board 2003, 2007
- NEMLAR: Network for Euro-Mediterranean Language Resources, International Conference on Arabic Language Resources and Tools, Scientific Committee, 2004
- ICSLP: International Conference on Spoken Language Processing: Scientific Review Committee 2002
- Penn Linguistics Colloquium: Reviewer

Advisory

- National Institute of Standards and Technology, Organization of Scientific Area Committees (OSAC): IT/Multimedia Scientific Area Committee, Speaker Recognition Subcommittee, 2014-present
- LINDAT-CLARIN Advisory Board: Czech/EU funded project on resource sharing for humanities research and language technology development, 2011-present
- Scientific Working Group for Forensic and Investigatory Speaker Recognition, 2013-2014
- COCOSDA: International Committee for the Coordination and Standardization of Speech Databases: North American Rapporteur, 2003-2012
- Hindi Treebank: NSF funded project: Scientific Advisory Board, 2009-2011
- WRITE: International Committee for Written Language Resources: North American Rapporteur, 2003-2010
- SILT: NSF funded project on Sustainable Infrastructure for Linguistic Technology: Scientific Advisory Board, 2009
- FlareNet: Fostering Language Resources Network, Individual Subscriber and Partner and Institutional Member Representative, 2008-
- OLAC: Open Language Archives Community: OLAC Council, 2003-
- ANC: American National Corpus: Steering Committee, 2002-

Other Committees

- Cambridge University Press: Book Proposal Referee
- Harcourt Brace: Reviewer
- Linguistic Data Consortium: Executive Director
- DARPA TIDES Program: Advisory Committee Alternate
- University of Pennsylvania, Computer Services Restructuring project: Steering Group
- University of Pennsylvania, Computer Services Benchmarking project: Task Force

Languages

Achieved proficiency in Italian, French, Portuguese
Also formally studied German, Latin, Polish, Irish Gaelic
Analyzed aspects of: Italian (phonology), Amazigh (morphology), Sranan (serial verb morphology and syntax), Takelma (phonology and morphology), K'ekchi (phonetics of rapid speech), Portuguese (verb morphology), Papiementu (verb morphology), Korean (verb morphology)

Publications and Presentations

- (Forthcoming) 2015, Data Bases and Statistical Systems: Linguistics in International Encyclopedia of Social & Behavioral Science 2nd Edition, James Wright, ed. Elsevier.
- January 2015, A License Scheme for a Global Federated Language Service Infrastructure (with Denise DiPersio), WLSI 2015: The Second International Workshop on Worldwide Language Service Infrastructure, Kyoto, January 22-23.

- January 2015, Incorporating Speech Technology into Language Archive Development (with Mark Liberman), Linguistic Society of America Annual Meeting, Symposium: Making the Most of Language Archives: Automatic Analysis of Audio and Video for Enhanced Linguistic Analysis, January 8-11, Portland, OR.
- November 2014, Special Issue on Archiving Sociolinguistic Data (issue editor with Malcah Yaeger-Dror), Language and Linguistics Compass: Sociolinguistics, 8:3.
- November 2014, Challenges and Opportunities in Sociolinguistic Data and Metadata Sharing, Language and Linguistics Compass: Sociolinguistics, 8:3.
- August 2014, Intellectual Property Rights Management with Web Service Grids (with Denise DiPersio), COLING 2014 - 25th International Conference on Computational Linguistics, Workshop on Open Infrastructures and Analysis Frameworks for HLT, Dublin.
- August 2014, Data Sets from Independent Sources (with Malcah Yaeger-Dror), presented at Methods in Dialectology XV special session on Panel Studies: Challenges, Food for Thought and Ways Forward.
- May 2014, New Directions for Language Resource Development and Distribution (with Denise DiPersio, Mark Liberman, Andrea Mazzucchi, Stephanie Strassel, Jonathan Wright), Proceeding of LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik.
- May 2014, Facing the Identification Problem in Language-Related Scientific Data Analysis (with Joseph Mariani, Gil Francopoulo, Patrick Paroubek, Marine Delaborde), Proceeding of LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik.
- May 2014, Developing a Framework for Describing Relations among Language Resources (with Penny Labropoulou, Maria Gavrilidou), Proceeding of LREC 2014, Reykjavik.
- May 2014, The Language Application Grid (with Nancy Ide, James Pustejovsky, Eric Nyberg, Denise DiPersio, Keith Suderman, Marc Verhagen Di Wang, Jonathan Wright), Proceeding of LREC 2014, Reykjavik.
- January 2014, Dimensions of Speaker Recognition Research Data, presented at LSA Symposium: Data for Empirical Foundations of Forensic Linguistics, Minneapolis (with Mark Liberman).
- December 2013, Data Center Models and Impact on Scientific Research Communities, presented at 1st IEEE Global Conference on Signal and Information Processing symposium on Advancing Neural Engineering Through Big Data, December 3-5, 2013 Austin, Texas.
- October 2013, Sharing, Structuring and Processing Data: Part 1: Advantages and Challenges, NWAV42: New Ways of Analyzing Variation, Pittsburgh.
- July 2013, From STrans to WebAnn: the Evolution of LDC Corpus Creation Tools, Séminaire DGA: Traitement de l'Information Multimédia, Paris.
- July 2013, Language Resources in Recent US Program, Séminaire DGA: Traitement de l'Information Multimédia, Paris.
- June 2012, New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus (with Stephanie Strassel, Kevin Walker, Karen Jones, Dave Graff), Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28.
- May 2012, Language Resources for Public Security Applications: a Data Center Perspective, Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation, Istanbul.
- May 2012 LDC Language Resource Papers Catalog: Building a Bibliographic Database (with Eleftheria Ahtaridis, Denise DiPersio), Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation, Istanbul.
- May 2012, Twenty Years of Language Resource Development and Distribution: A Progress Report on LDC Activities (with Mark Liberman), Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation, Istanbul.
- May 2012, Toward the Harmonization of Metadata Practice for Spoken Languages Resources (with Malcah Yaeger-Dror), SpeechCorpora 2012: Workshop on Best Practices for Speech Corpora in

Linguistic Research, LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27.

- 2011, Technical Infrastructure Supporting Large-Scale Linguistic Resource Creation (with Kazuaki Maeda, Andrea Mazzucchi), in Olive, Joseph, Caitlin Christianson and John McCary (eds.) Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, New York, Springer.
- November 2010. Robust, Digital, Empirical, Reproducible, Sociolinguistic, Methodology (with Stephanie Strassel), in Towards Best Practices in Sociophonetics, Workshop at NWAV39: New Ways of Analyzing Variation, San Antonio, Texas, November 4-6, 2010.
- October 2010, Making a Field Recording, Chapter 3 in Di Paolo, Marianna and Malcah Yeager-Dror, 2010, Sociophonetics: A Student's Guide, Routledge.
- May 2010, Adapting to Trends in Language Resource Development: A Progress Report on LDC Activities (with Mark Liberman), LREC 2010: Seventh International Language Resources and Evaluation Conference, Valletta, Malta.
- May 2010, A Road Map for Interoperable Language Resource Metadata (with Khalid Choukri, Nicoletta Calzolari, D. Terence Langendoen, Johannes Leveling, Martha Palmer, Nancy Ide, James Pustejovsky), LREC2010: Seventh International Language Resources and Evaluation Conference, Valletta, Malta.
- May 2010, Greybeard – Voice and Aging (with Linda Brandschain, David Graff, Kevin Walker, Chris Caruso and Abby Neely), LREC 2010: Seventh International Language Resources and Evaluation Conference, Valletta, Malta.
- May 2010, The Mixer 6 Corpus: Resources for Cross-Channel and Text Independent Speaker Recognition (with Linda Brandschain, David Graff, Kevin Walker, Chris Caruso and Abby Neely), LREC2010: Seventh International Language Resources and Evaluation Conference, Valletta, Malta.
- February 2010, From Road Maps to Plans: Towards the Design and Cost-Benefit Analysis of a Universal Language Resource Catalog (with Khalid Choukhri), invited paper, FLaReNet Forum 2010: Language Resources of the Future - the Future of Language Resources, Institut d'Estudis Catalans, Barcelona, Spain.
- January 2010, Interoperability in the Context of Large Data Centers and Common Task Programs, invited paper at ICGL 2010: 2nd International Conference on Global Interoperability for Language Resources, City University of Hong Kong, January 18-20.
- October 2009, Models of Phonological Variation for Multi-dialectal Communities: the case of L'Aquila, NWAV 38: The 38th Annual Meeting of New Ways of Analyzing Variation, University of Ottawa, Ottawa, Canada.
- October 2009, Closer Still to a Robust, All Digital, Empirical, Reproducible Sociolinguistic Methodology (with Stephanie Strassel), NWAV 38: The 38th Annual Meeting of New Ways of Analyzing Variation, University of Ottawa, Ottawa, Canada.
- September 2009, The Broadcast Narrow Band Speech Corpus: A New Resource Type for Large Scale Language Recognition (with Linda Brandschain, Abby Neely, David Graff, Kevin Walker, Chris Caruso, Alvin Martin, Craig Greenberg), Interspeech 2009, Brighton, UK.
- August 2009, Basic Language Resources for Diverse Asian Languages: A Streamlined Approach for Resource Creation (with Heather Simpson, Kazuaki Maeda), In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore.
- June 2008, The Linguistic Data Consortium Member Survey: Purpose, Execution and Results (with Marian Reed, Denise DiPersio), LREC2008: Sixth International Language Resources and Evaluation Conference, Marrakesh, Morocco.

- June 2008, 15 Years of Language Resource Creation and Sharing: A Progress Report on LDC Activities (with Mark Liberman), LREC2008: Sixth International Language Resources and Evaluation Conference, Marrakesh, Morocco.
- June 2008, Bridging the Gap between Linguists and Technology Developers: Large-Scale, Sociolinguistic Annotation for Dialect and Speaker Recognition (with Stephanie Strassel, Meghan Glenn, Reva Schwartz, Wade Shen, Joseph Campbell), LREC2008: Sixth International Language Resources and Evaluation Conference, Marrakesh, Morocco.
- June 2008, Speaker Recognition: Building the Mixer 4 and 5 Corpora (with Linda Brandschain, David Graff, Abby Neely, Kevin Walker), LREC 2008: Sixth International Language Resources and Evaluation Conference, Marrakesh, Morocco.
- June 2008, Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources (with Heather Simpson, Kazuaki Maeda, Kathryn Baker, Boyan Onyshkevych), LREC 2008: Sixth International Language Resources and Evaluation Conference, Marrakesh, Morocco.
- October 2007, Phonological Variation in Multi-Dialectal Italy: distinguishing e from ε, NWAV 2007, Philadelphia.
- September, 2007 Linguistic Resources in Support of Various Evaluation Metrics (with Stephanie Strassel, Meghan Lammie Glenn, Lauren Friedman), MT Summit XI, Workshop on Automatic Procedures in MT Evaluation, Copenhagen.
- August 2007, Towards an Integrated Understanding of Speech Overlaps in Conversation (with Jiahong Yuan, Mark Liberman), ICPhS XVI: The International Congress of Phonetic Sciences, Saarbrücken, August 6-10.
- August 2007, Construction of a Phonotactic Dialect Corpus using Semiautomatic Annotation (with Reva Schwartz, Wade Shen, Joseph Campbell, Shelley Paget, Julie Vonwiller, Dominique Estival), Interspeech 2007: Eighth Annual Conference of the International Speech Communication Association, Antwerp.
- August 2007, Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora (with Linda Corson, David Graff, Kevin Walker), Interspeech 2007: Eighth Annual Conference of the International Speech Communication Association, Antwerp.
- April 2007, Linguistic Resources and the Development and Evaluation of Text and Speech Systems, Chapter 8 of Dybkjær, Laila, Holmer Hemsen, Wolfgang Minker, Evaluation of Text and Speech Systems, Springer Netherlands.
- December 2006, HKUST/MTS: A Very Large Scale Mandarin Telephone Speech Corpus (with Yi Liu, Pascale Fung, Yongsheng Yang, Shudong Huang, David Graff), ISCSLP 2006: Proceeding of the 5th International Symposium on Chinese Spoken Language Processing, Singapore, December 13-16, 2006.
- September 2006, Towards an Integrated Understanding of Speaking Rate in Conversation (with Jiahong Yuan, Mark Liberman), Interspeech 2006 – ICSLP: Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21.
- April 2006, Linguistic Data Resources (with Mark Liberman, Victoria Arranz, Khalid Choukri), Chapter 3 of Schultz, Tanja and Katrin Kirchoff, eds., Multilingual Speech Processing, Elsevier, Academic Press.
- May 2006, Integrated Linguistic Resources for Language Exploitation Technologies (with Stephanie Strassel, Andrew Cole, Denise DiPersio, Mark Liberman, Xiaoyi Ma, Mohamed Maamouri), Language Resources and Evaluation Conference, Genoa, Italy.
- May 2006, The Mixer and Transcript Reading Corpora: Resources for Multilingual, Cross-channel Speaker Recognition Research (with Walt Andrews, Joseph P. Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, Kevin Walker), Language Resources and Evaluation Conference, Genoa, Italy.

- May 2006, More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities (with Mark Liberman), Language Resources and Evaluation Conference, Genoa, Italy.
- May 2006, What is Quality?, invited talk at the Workshop on Quality Assurance and Quality Measurement for Language and Speech Resources, LREC 2006: 5th International Conference on Language Resources and Evaluation, Genoa, May 22-28
- May 2006, Customized Corpora for Speech Controlled Computing (with Kazuaki Maeda), Language Resources and Evaluation Conference, Genoa, Italy.
- May 2006, Corpus Support for Machine Translation at LDC (with Xiaoyi Ma), Language Resources and Evaluation Conference, Genoa, Italy.
- May 2006, Low-cost Customized Speech Corpus Creation for Speech Technology Applications (with
- Kazuaki Maeda, Kevin Walker), Language Resources and Evaluation Conference, Genoa, Italy.
- December 2005, Evaluation Initiatives: US Activities in HLT Evaluation, European Language Resource Association, Human Language Technology Evaluation Workshop, Malta.
- September 2005, Building Corpora for Multiple Speaker Identification Scenarios, ITIC Speaker Identification Workshop, McLean, VA.
- May 2005, DARPA-Sponsored Research in Human Language Technology: The Role of Shared Data (with Jack Godfrey, Mark Liberman, Charles Wayne), International Conference on Intelligence Analysis: Methods and Tools, McLean, VA.
- December 2004, Development of a Chinese Telephony Conversational Corpus for Speech Processing (with Liu Yi, P. Fung, S. Huang, Z. Lufeng, C. Benfeng), International Symposium on Chinese Spoken Language Processing.
- December 2004, Dialectal Arabic Orthography-Based Transcriptions and CTS Levantine Arabic Collection (with Mohamed Maamouri, Dave Graff, Hubert Jin, Tim Buckwalter), EARS/RT-04 Workshop.
- September 2004, Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions (with Mohamed Maamouri, Tim Buckwalter), presented at the NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Sept. 22-23, 2004 .
- June 2004, The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation (with Joseph P. Campbell, Hirotaka Nakasone, David Miller, Kevin Walker, Alvin Martin, Mark Przybocki), Speaker Recognition Odyssey, Toledo.
- May 2004, The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text (with David Miller, Kevin Walker), Language Resources and Evaluation Conference, Lisbon, Portugal.
- May 2004, The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data (with Joseph P. Campbell, Hirotaka Nakasone, David Miller, Kevin Walker), LREC 2004: Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon, Portugal.
- May 2004, TalkBank: Building an Open Unified Multimodal Database of Communicative Interaction (with Brian Macwhinney, Steven Bird, Craig Martell), LREC 2004: Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon, Portugal.
- May 2004, Progress Report from the Linguistic Data Consortium: recent activities in resource creation and distribution and the development of tools and standards (with Mark Liberman), LREC 2004: Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon, Portugal.
- May 2004, Collaborative Commentary, LREC 2004: Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon, Portugal.
- April 2004, Roadmaps and Resource Maps: Coordinating Language Resources for Cooperative Technology Development & Evaluation (with Kazuaki Maeda), COCOSDA/ICWLR Workshop, International Congress on Acoustics, Kyoto.
- October 2003, Robust Sociolinguistic Methodology: Tools, Data and Best Practices (with Stepanie Strassel), a workshop presented at NWAVE 32, the 32nd conference on New Ways of Analyzing Variation, University of Pennsylvania, Philadelphia, PA.

- September 2003, From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields (with David Miller, Kevin Walker) 8th European Conference on Speech Communication and Technology, Eurospeech 2003 - Interspeech 2003, Geneva, Switzerland, September 1-4.
- September 2003, Shared Resources for Robust Speech-to-Text Technology (with Stephanie Strassel, David Miller, Kevin Walker), 8th European Conference on Speech Communication and Technology, Eurospeech 2003 - Interspeech 2003, Geneva, Switzerland, September 1-4.
- August 2003, Corpora and Human Language Technologies for Language Learners (with J. Scott Payne, Kathleen Egan), EuroCALL 2003, Limerick, Ireland.
- August 2003, Core Linguistic Resources for the World's Languages (with Mike Maxwell, Stepanie Strassel), Joint Meeting of ELSNET: The European Network of Excellence in Human Language Technologies and ENABLER: The European National Activities for Basic Language Resources Network, Paris, France.
- March 2003, Linguistic Resource Creation for Research and Technology Development: A Recent Experiment (with Stephanie Strassel and Mike Maxwell), Association for Computing Machinery Transactions on Asian Language Information Processing, TALIP Volume 2, Issue 2, pp. 101–117.
- October 2002, Sharable Resources for Sociolinguistic Research (with Stephanie Strassel and William Labov), Proceedings of the 31st Conference on New Ways of Analyzing Variation, Stanford University, CA.
- June 2002, The DASL Project: a Case Study in Data Re-Annotation and Re-Use (with Stephanie Strassel), LREC 2002: Proceedings of the Third International Language Resource and Evaluation Conference, Las Palmas, Spain.
- June 2002, TIDES Language Resources: A Resource Map for Translingual Information Access, (with Mark Liberman), LREC 2002: Proceedings of the Third International Language Resource and Evaluation Conference, Las Palmas, Spain.
- June 2002, Language Resource Creation and Distribution at the Linguistic Data Consortium: A Progress Report (with Mark Liberman), LREC 2002: Proceedings of the Third International Language Resource and Evaluation Conference, Las Palmas, Spain.
- April 2002, Resources for Arabic Natural Language Processing (with Mohamed Maamouri), International Symposium on Processing Arabic, Tunis, Tunisia.
- March 2002, Christopher Cieri, David Miller, Kevin Walker, Research Methodologies, Observations and Outcomes in (Conversational) Speech Data Collection, HLT 2002 The Human Language Technologies Conference, San Diego, CA.
- 2002, Corpora for Topic Detection and Tracking (with David Graff, Nii Martey, Stephanie Strassel and Mark Liberman) in Allan, James, Jaime Carbonell, Jon Yamron, Topic Detection and Tracking Research, Kluwer International Series on Information Retrieval, W. Bruce Croft, series editor.
- November, 2001, Resources and Infrastructure to Support Robust, Omnipresent ASR (with David Graff, David Miller, Kevin Walker), Communicator, SPINE, ROAR Workshop, Orlando.
- November 2001, SPINE 2001 Data Preparation and Annotation and the SPINE Corpora (with Andy Cole, Dave Graff, Nii Martey, Stephanie Strassel, Cristina Tofan) Communicator, SPINE, ROAR Workshop, Orlando, FL.
- July 2001, Annotation Graphs, Annotation Servers and Multi-Modal Resources: Infrastructure for Interdisciplinary Education (with Steven Bird), Research and Development, Proceedings of the Association for Computational Linguistics: Workshop on Sharing Tools & Resources, Toulouse.
- March 2001, Switchboard Cellular Resources for Speaker Recognition (with David Miller and Kevin Walker), Speaker Recognition Workshop, Maritime Institute of Technology and Graduate Studies, Linthicum MD.
- March 2001, Shared Resources and Community Building for Corpus Linguistics and Language Teaching (with Steven Bird and Stephanie Strassel), The Third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA.

- March 2001, Getting SMART about Authoring (with Lea Christiansen, Kathleen Egan, Anita Kulman, Milton Paul), Presented at CALICO 2001: Computer-Aided Language Instruction Consortium, University of Central Florida, Orlando.
- March 2001, Data and Annotations for Sociolinguistics: A Corpus-Based Approach to Sociolinguistic Research (with Stephanie Strassel), The 25th Annual Penn Linguistics Colloquium, Philadelphia, PA.
- October 2000, I'll have a shot of noisy speech and put it in a dirty glass: The SPINE1 Training and Evaluation Corpora (with Kara Rennert and Stephanie Strassel), Proceedings of the Speech in Noisy Environments Workshop, Naval Research Laboratories.
- October 2000, Resources for Robust Analyses of Natural Language, Proceedings of the ROMAND 2000: Workshop on Robust Methods in Analysis of Natural Language Data, Lausanne, Switzerland.
- June 2000, Issues in Corpus Creation and Distribution: the Evolution of the Linguistic Data Consortium (with Mark Liberman), Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece.
- June 2000, Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts (with Dave Graff, Mark Liberman, Nii Martey and Stephanie Strassel), Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece.
- June 2000, Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora (with Strassel, Stephanie, Dave Graff, Nii Martey), Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece.
- May 2000, Resources, New and Forthcoming, from LDC (with Dave Graff, Stephanie Strassel), Speech Transcription Workshop, University of Maryland, May 16-19, 2000.
- March 2000, Multiple Annotations of Reusable Data Resources: Corpora for Topic Detection and Tracking, Rajman, M. and J. C. Chappelier, eds., Actes des 5es Journees internationales d'analyse statistique des donnees textuelles, volume 1, Ecole Polytechnique Federale de Lausanne.
- February 2000, The TDT-3 Text and Speech Corpus (with David Graff, Nii Martey, Stephanie Strassel), Topic Detection and Tracking Workshop, Vienna, Virginia.
- January 2000 Issues and Tools for Annotating a Corpus of Sociolinguistic Field Data, Linguistic Exploration Workshop in conjunction with the Linguistic Society of American Annual Meeting, Chicago.
- October 1999, Corpus Sociolinguistics: Issues, Data and Tools (with Stephanie Strassel), NWAVE-28, York University, Toronto, Ontario.
- June 1999, This Ain't Your Father's Digital Data: Another Perspective on Legal Information, CALI 1999 - The Conference for Law School Computing. Eugene Oregon.
- June 1999, Telephone Speech Corpora: New Needs, Languages, Methods and Technology (with Alexandra Canavan, Kevin Walker, David Graff), Hub-5 Conversational Speech Understanding (LVCSR) Workshop, Maritime Institute Technology and Graduate Studies, Linthicum Heights, Maryland.
- February 1999, The TDT-2 Text and Speech Corpus (with David Graff, Mark Liberman, Nii Martey, Stephanie Strassel), DARPA Broadcast News Workshop, Washington, DC.
- November 1998, Topic Detection and Tracking Corpora (with Dave Graff), TREC/SDR Conference, Gaithesburg Maryland.
- September 1998, Update on Lexical Resources and Projects at the Linguistic Data Consortium (with Dave Graff), Ninth Hub-5 Conversational Speech Recognition (LVCSR) Workshop, Maritime Institute Technology and Graduate Studies, Linthicum Heights, Maryland.
- May 1998, The Creation, Distribution and Use of Linguistic Data: the Case of the Linguistic Data Consortium (with Mark Liberman), In Proceedings, First International Conference on Language Resources and Evaluation, Granada, Spain.
- June 1997, Panda's Don't Have Thumbs-and other evolutionary lessons from Penn Law's digital library project (with Don Arndt), CALI-LEAP Conference, IIT Chicago Kent Law School, Chicago.

- June 1995, Treasure Trove or Black Hole: The Internet and Language Studies, invited speaker Corso di Perfezionamento in Lingue Moderne, University of Rome "La Sapienza", Rome, Italy.
- May 1995, Ethnic and Geographic Variation in a Language Contact Situation, invited speaker Corso di Perfezionamento in Lingue Moderne, University of Rome "La Sapienza", Rome, Italy.
- February 1993, Some Recent Advancements in the Probabilistic Determination of Linguistic Relationships, Penn Linguistics Colloquium, Philadelphia, PA.
- February 1991, Italian Americans: Linguistic Portraits, Penn Linguistics Colloquium, Philadelphia, PA.
- October 1989, Phonological Variation in a Language Contact Situation - the interaction of ethnicity and geographic space (with Cheryl Cerrito), NWAVE XVIII/ADS-C, Duke University, Durham, NC.
- August 1989, Statistical Methods to Determine Multi-Lingual Computer Text Files (with John Fought, George Milliken and Jim Waters), 100th Anniversary meeting of the American Statistical Association, Washington, DC.
- January 1989, The Phonological Toolbox, Penn Linguistics Colloquium, Philadelphia, PA.
- April 1988, Technology and the Teaching of Foreign Language in Primary School, invited speaker at the Congress on Foreign Language Teaching in Primary Education, University of L'Aquila, L'Aquila, Italy.
- December 1986, The Application of Microcomputer Technology to Language Learning and Language Teaching - sessione pratica, featured presentation in "Il Computer nella didattica delle lingue", University of L'Aquila, L'Aquila, Italy.
- December 1986, The Application of Microcomputer Technology to Language Learning and Language Teaching - sessione teoretica, featured presentation in "Il Computer nella didattica delle lingue", University of L'Aquila, L'Aquila, Italy.
- November 1986, The Role of the Computer in Second Language Acquisition Research (with Dana Boatman and Nadine O'Connor), ADCIS - 28th International Conference, Arlington, Virginia.
- November 1986, The Computer's Contribution to Solving some Problems in Second Language Reading Tasks (with Dana Boatman and Nadine O'Connor), ADCIS - 28th International Conference, Arlington, Virginia.
- May 1986, Computerized Second Language Acquisition Research (with Dana Boatman and Nadine O'Connor), CALICO, Naval Academy, Annapolis, Maryland.
- May 1986, The Linguist's Toolbox: a progress report on a software package for linguists (co-author John Fought), CALICO, Naval Academy, Annapolis, Maryland.
- April 1986, The Role of the PC in Researching Learning Behavior (with Dana Boatman and Nadine O'Connor), Penn Review of Linguistics, Spring 1986.
- January 1986, In Search of the Origin of Italo-American Loanword Phonology, Penn Linguistics Colloquium, University of Pennsylvania, Philadelphia, Pennsylvania.
- January 1986, The PC-DOS Environment and Developing Utilities for Linguists, Penn Linguistics Colloquium PC Workshop, University of Pennsylvania, Philadelphia, Pennsylvania.
- January 1986, The Role of the PC in Researching Learning Behavior (with Dana Boatman and Nadine O'Connor), Penn Linguistics Colloquium PC Workshop, University of Pennsylvania, Philadelphia, Pennsylvania.
- October 1985, Borrowing and Loanword Assimilation in the Italo-American Community, N-WAVE XIV, Georgetown University, Washington, DC.
- September 1985, "And" and the English Narrative: A Story of Continuity, Penn Review of Linguistics, Spring 1985.
- January 1985, "And" and the English Narrative: A Story of Continuity, Penn Linguistics Colloquium, Philadelphia, Pennsylvania.
- October 1984, The Optical Scanner and its Position in the Creation of Textual Data Bases, N-WAVE XIII Text Analysis Workshop, University of Pennsylvania, Philadelphia, Pennsylvania.

Employment History

1/98 – Present Linguistic Data Consortium, Philadelphia, PA
Position: Executive Director. Provide overall leadership for Linguistic Data Consortium. Establish accountability for the organization and its unique business model. Exercise overall responsibility for planning, operations, external relations and financial performance.

1/90 - 12/97 University of Pennsylvania Law School, Philadelphia, PA
Position: Director of Computer Services. Hired as Information Management Specialist; promoted six levels to Director within my first three years. Responsible for planning and budgeting of all Law School computer operations including supervision of 5 full-time, 4 part-time and 20+ student staff members. Responsible for representing the Law School on University task forces (Restructuring Computing across the University Committee, Information Systems and Computing Steering Committee, Provost's Year 2000 Committee, Provost's WWW Committee, Law School Senior Staff, Law Library Steering Council, Liaison to Small Schools Committee) and at professional conferences. Rebuilt Law School computing infrastructure. Reorganized Computer Services department and revised service portfolio. Created Professional Development, Cross Training and Zero-Downtime programs for computing staff. Introduced Computing Help Desk and universal, standard access to supported services for faculty, staff and students. Provided technical expertise for planning of the new Tanenbaum Hall facility and implemented novel approaches to networking in a virtual lab environment. Co-wrote grant proposal for, developed and managed the Digital Library Text/Imaging Initiative.

12/88 - 12/89 University of Pennsylvania Language Analysis Center, Philadelphia, PA
Position: Research Specialist. Designed and managed the development of the Phonological Toolbox, a software package to help phonologists complete analyses of transcribed text. Performed marketing functions including demonstration of products and creation of proposals. Consulted on the creation of 4 online bilingual dictionaries and the reorganization of a 15 million word multilingual text archive. Trained and supervised student workers. Developed and implemented a training program to teach linguists to use and program within Unix and DOS environments. Administered a PC cluster.

9/85 - 12/88 University of Pennsylvania Language Analysis Project, Philadelphia, PA
Position: Research Assistant. Responsibilities include all aspects of the construction of program products for linguistic analysis and language pedagogy (high and low level design, coding, testing, implementation, optimization, maintenance, documentation, use, evaluation and review); developing and maintaining language specific data tables and an 15 million word text base spanning 35 "critical" languages; porting software tools from Apollo Domain UNIX to MS DOS; evaluating PC networks in Computer Aided Language Learning; administering a PC lab (including hardware and software installation, evaluation and development, management of network links between PCs and Apollo systems); creating local documentation and training new users.

9/83 - 8/85 University of Pennsylvania Language Analysis Project, Philadelphia, PA
Position: Scanning Operator. Responsibilities included all aspects of the operation of the Kurzweil Data Entry Machine (KDEM 3000): creating documentation, developing training sets and inputting documents; the development of an environment for building multilingual text bases: input, editing, developing screen and printer fonts, training new users and creating documentation and the development of text analytical tools.

Consulting

8/93 - 3/95 New England Nuclear, Boston, MA

Project: Database to Catalog Conversion. Subcontractor for DuPont NEN Division. Designed, wrote, maintained and updated software to generate catalog, price list and product list from proprietary database format. Data is first parsed, converted into SDF for incorporation into dBase database for sorting and manipulation and then into RTF for layout and publication.

4/88 - 6/88 Università dell'Aquila Istituto Inglese, L'Aquila, Italy
Position: Researcher/Lecturer. Taught seminars entitled The Syntax of English Poetry and The Pronunciation of American English, acted as a substitute teacher for classes in advanced English, developed a complete concordance of the Edward Lear's Book of Nonsense, helped develop software to analyze rhyme and meter in English poetry and lead a conference on the computer and foreign language teaching in primary school.

7/87 - 12/88 E. I. Du Pont de Nemours Inc. Technical Publications, Wilmington, DE
Project: Systems Upgrade and Networking. Headed the Systems Group which evaluated, planned, justified, purchased, installed, configured, maintained, serviced, administrated, documented and developed training for computer systems in a Du Pont corporate publishing group. Was responsible for the function of approximately 35 IBM PCs, 25 Apple Macintoshes, 5 Apollo workstations, a CCI typesetting system, 20 laser printers, 15 dot matrix printers and various peripherals. Provided these computer resources for approximately 70 writers, illustrators, artists, typesetters, compositors, coordinators, supervisors, managers and support staff. Managed a spending budget of approximately $250K/annum. Also functioned as liaison to corporate MIS groups and consulted on new approaches to computer based publishing, multi-platform networking and cross platform compatibility.

10/86 - 12/86 Università dell'Aquila Centro Linguistico Interdipartimentale L'Aquila, Italy
Project: Laboratory for Computer Assisted Language Studies. Recruited as specialist computer assisted language studies (language learning and teaching, linguistic analysis and literary research). Responsibilities included all aspects of the establishment of a CALS laboratory (needs analysis, initial planning and projections, policy administration, initial hardware and software purchases, training and daily maintenance), the acquisition of hardware and software, the development of novel materials including courseware and courseware development packages, the training of staff and students, the development of new courses in English as a Second Language (ESL) using CALS materials, and the public presentation of progress reports on the development efforts at the center. Based on my performance, I was offered an "open contract" at the center.

6/86 - 10/86 John G. Fought and Associates, Philadelphia, PA
Project: Multilingual Office Products Market Analysis. Contracted to Xerox Multilingual Products Division to analyze the market for multilingual office products. Responsibilities included estimation of foreign market potential, identification and evaluation of competing multilingual systems, including hardware and software, and the investigation of potential co-development opportunities.

8/86 - 10/86 University of Pennsylvania Audio-Visual Center, Philadelphia, PA
Project: Audio Authoring Template. Developed a computerized lesson authoring template that integrated PCs and Tandberg computer controlled cassette recorders and allowed authors to associate textual and auditory stimuli in lessons as well as collect, record and evaluate student responses.

Continuing Education

- February 2013, Collaborative Institutional Training Initiative (CITI) Human Research Curriculum, Social/Behavioral Research Course

- August 2003, Project Management - one-day Franklin Covey workshop hosted by the Linguistic Data Consortium
- March 2000, Time Management - one-day Franklin Covey workshop hosted by the University of Pennsylvania on principles of time management, setting priorities and planning activities.
- October 1997, Managing Digital Imaging Projects - three day workshop offered by the Research Libraries group and the Library of Congress covering planning and estimating, hardware and software, outsourcing, process and quality assurance for digital imaging projects.
- August 1996, Leadership in Times of Change - five day workshop offered to senior managers at the University of Pennsylvania covering personality type assessment, change hardiness and change management.
- June 1995, Developing HTML Pages - one day workshop offered at the annual CALI-LEAP conference for computing professionals in Law covering tools and techniques for using HTML to create worldwide web pages.

Software Development Projects in Linguistics (through 2005)

I did software development regularly through 2005 but not much since. The list is in chronological order, most recent first.

- Quantitative/Acoustic Analysis and Tools – toolkit based on Entropics utilities, Perl and TK widgets that assists linguist in searching and coding a set of time aligned transcripts. Possible analyses include wide and narrow band spectrograms, formant, pitch and power tracks.
- Ringe/Oswalt Statistical Methods Tool - applies multiple algorithms to comparative language data to determine the likelihood of a historical relationship between the languages.
- Korean Verb Processor - morphological processor to analyze verb forms and generate paradigms for both regular and irregular stems.
- Phonological Toolbox - multilingual, theory independent software application for the phonological analysis of transcribed data.
- Linguist's Toolbox - integrated software applications providing computational support for multilingual text linguistics.
- English Verb Generator - morphological processor to build regular and irregular verb paradigms.
- Segmentation Manager - integrated utilities to guide a linguist through a morphological analysis, managing intermediate stages of data and providing feedback at each stage.
- Bilingual Dictionaries Project - electronic, SGML-tagged, bilingual dictionaries for Arabic, Spanish and Korean.
- Old French Reader - text browser for Old French with integrated morphological processor and lexical data base.
- English Reader - text browser and lexical database that can be configured for students from different first language backgrounds.
- Lear Concordance – edited complete lexical concordance of Edward Lear's Book of Nonsense.
- Text Works - 15 million word text database spanning 35 "critical" languages.
- French Tense Cloze Template - testing software to present students with questions on French Past Tense usage, collects answers, times responses, analyzes results and prepares reports to the teacher/researcher.
- TestGen - utility to generate unique language exams conforming to a teacher-specified template using questions selected at random from a database.

Synergistic Activities

- As Executive Director of the Linguistic Data Consortium, I oversee the creation and distribution of language data, tools and standards that benefit more than 3500 organizations in 71 countries.

- As a member of the Organizing Committee for the EMELD (Electronic Metastructure for Endangered Languages Data) project's, Digital Tools Summit in Linguistics, I helped plan and execute two workshops (2006, 2007) that brought together field linguists, technology developers and funders to stimulate new collaborations in the areas of language documentation and preservation.
- As principal investigator of the project "Advance Linguistic Resources and Infrastructure Supporting Human Language Technology Development", I have overseen the creation of more than 35 different corpora used in the DARPA TIDES (Translingual Information Detection Extraction and Summarization) and EARS (Effective Affordable, Reusable Speech-to-Text) program. These corpora include billion-word news text archives, Treebanks, collections of conversational telephone speech involving more than 15,000 subjects. Several of these corpora have already been released for general research use; all of the corpora will be published within the next two years.
- As a member of Council of the Open Language Archives Community, I have contributed to the specification and implementation of a system that archives metadata describing hundreds of speech and text resources created by more than 20 organizations making that metadata available to researchers worldwide in a single location.
- As the North American Rapporteur for both COCOSDA and ICWLRE, international committees for the coordination of speech and text databases, I produce and publish a report that describes funding and activities in North America focused on resource creation and distribution.
- As a member of the American National Corpus (ANC) Steering Committee, I coordinated several data contributions to the ANC, drafted the user agreements for data providers and users, helped plan the organization of the ANC Consortium and oversaw the first release of the corpus.

Collaborators & Other Affiliations

- William Labov (U. Penn) is my thesis advisor. Other readers were Mark Liberman (LDC, UPenn) and Francesco Avolio (U. L'Aquila)
- Co-Editors: Malcah Yaeger-Dror (U. Arizona, 2014)
- Editors: James Allan (U. Mass, 2002), Jaime Carbonell (CMU 2002), Marianna DiPaolo (U. Utah, 2010), Laila Dybkjær (U. Southern Denmark, NISLab, 2007), Holmer Hemsen (DFKI, 2007), Katrin Kirchoff (U. Washington, 2006), Wolfgang Minker (U. Ulm, 2007), Doug Oard (UMBC), Tanja Schultz (CMU, 2006), Malcah Yaeger-Dror (U. Arizona), Jon Yamron (Dragon Systems 2002) were editors of volumes to which I submitted chapters.
- Co-authors at Penn include: Eleftheria Ahtaridis (2014), Alexandra Canavan (1999), Cheryl Cerrito (1989), Andy Cole (2006), Linda Brandschain (née Corson, 2010), ChrisCaruso (2010), Denise DiPersio (2015), John Fought (1989), Lauren Friedman (2007), Meghan Glenn (née Lammie, 2008), David Graff (2012), Shudong Huang (2006), Hubert Jin (2004), Karen Jones (2012), William Labov (2002), Mark Liberman (2014), Xiaoyi Ma (2006), Mohamed Maamouri (2006), Kazuaki Maeda (2011), Nii Martey (2002), Andrea Mazzucchi (2014), David Miller (2004), Abby Neeley (2010), Marian Reed (2008), Kara Rennert (2000), Heather Simpson (2009), Stephanie Strassel (2014), Cristina Tofan (2001), Kevin Walker (2012), Jonathan Wright (2014), Jiahong Yuan (2007)
- Other co-authors include: Walt Andrews (USG, 2006), Don Arndt (Nebraska, 1997), Victoria Arranz (ELDA, 2006), Kathryn Baker (USG 2008), Chen Benfeng (HKUST, 2007), Steven Bird (U. Melbourne, 2004), Dana Boatman (JHU, 1986), Tim Buckwalter (UMD CASL, 2004), Nicoletta Calzolari (ILC-CNR, 2010), Joseph P. Campbell (MIT-LL, 2008), Khalid Choukri (ELDA, 2010), Lea Christiansen (USG, 2002), Marine Delaborde (LIMSI-CNRS, 2014), George Doddington (SRI, 2006), Kathleen Egan (National Virtual Translation Center, 2003), Dominique Estival (Appen 2007), Gil Francopoulo (Tagmatica, IMMI-CNRS, 2014), Pascale Fung (HKUST, 2007), Jack Godfrey (USG, 2006), Craig Greenberg (NIST, 2009), Nancy Ide (Vassar, 2014), Anita Kulman (USG, 2002), D. Terence Langendoen (NSF, Arizona, 2010), Johannes Leveling (DCU, 2010), Zhai Lufeng (HKUST, 2007), Joseph Mariani (LIMSI-CNRS & IMMI, 2014), Craig Martell (US NPS,

2004), Alvin Martin (NIST, 2009), Mike Maxwell (UMD CASL, 2003), George Milliken (Kansas State, 1989), Hirotaka Nakasone (FBI, 2006), Eric Nyberg (CMU, 2014), Nadine O'Connor (U. Chicago, 1986), Boyan Onyshkevych (USG, 2008), Shelley Paget (Appen, 2007), Martha Palmer (Colorado, 2010), Patrick Paroubek (LIMSI-CNRS, 2014), Milton Paul (USG, 2002), J. Scott Payne (Penn State, 2003), Mark Przybocki (NIST, 2006), Reva Schwartz (USG 2008), Wade Shen (MIT-LL, 2008), Keith Suderman (Vassar, 2014), Marc Verhagen (Brandeis, 2014), Julie Vonwiller (Appen 2007), Di Wang (CMU, 2014), Jim Waters (USG, 1989), Charles Wayne (USG 2005), Malcah Yaeger-Dror (U. Arizona, 2014), Yongsheng Yang (HKUST, 2007), Yi Liu (HKUST, 2007)