David E. Graff
803 S. 49th St.
Philadelphia, PA 19143
(215) 724-0640

CURRICULUM VITAE

EDUCATION:

- Recieved B.A. cum laude in linguistics from Pitzer College, Claremont, CA, in Spring of 1979. Coursework covered all major branches of linguistics, and included a semester in Paris (courses conducted in French) plus graduate-level courses at the 1978 Summer Institute of the Linguistic Society of America (Univ. of Illinois, Champagne-Urbana).

- Enrolled in the doctoral program in linguistics at the University of Pennsylvania from September 1979 through December 1995. (ABD)

WORK EXPERIENCE:

- July 1992 to present: Programmer/analyst for the Linguistic Data Consortium at the University of Pennsylvania. Duties include:

  - Provide software and technical support for LDC staff
  - Prepare corpora of text and speech data for publication
  - Develop, maintain and apply software as needed to process text and speech data bases for quality control, standardization, and enhancement for research uses
  - Plan, implement and manage the technical aspects of new speech corpus collection projects, recording from telephone and broadcast channels
  - Design, implement and maintain relational databases for the storage, annotation and publication of multi-lingual corpora and lexicons (with primary focus on Arabic), drawing from text and speech domains, with full linkage between lexicon and corpus text
  - Assist in the acquisition and configuration of computer resources, to support expanding needs for storage and computational capacity

  Projects and accomplishments include the following:

  - Assembly, conditioning, normalization and documentation for "Gigaword" text corpora in English, Arabic, Chinese, Spanish, and French, with corpus sizes ranging from 0.6 to 1.7 billion words per language.
  - Creation of database schemas and annotation tools for lexicon development based on speech and text corpora in both Modern Standard and dialectal Arabic, with full relational linkage between lexical morphology analysis data and individual tokens in corpus texts.

- Technical management and tool development for a relational database to control and document speech data collections from both telephone and interview conversational settings, involving tens of thousands of human subjects and recordings.

- Managed the creation and publication of a broadcast news speech corpus for the ARPA Continuous Speech Recognition Program (CSR); each of two project phases involved recording, transcribing and publishing 120 hours of television and radio news broadcasts; supervised or coordinated the efforts of 4 full-time and 12 part-time workers.

- Managed and performed technical aspects for the collection, annotation and publication of multi-month, multi-media news corpora to support information retrieval research under the DARPA project Topic Detection and Tracking (TDT); corpus creation involved daily recordings from radio and television networks (totaling over 1600 hours of broadcast audio) and from newswires, plus creation and time-marking of audio transcripts, and extensive topic annotations by dozens of annotators.

- November 1983 to July 1992: Member, engineering staff at RCA/GE Advanced Technology Laboratories, Moorestown, NJ. Major projects included the following:

  - Developed and applied software tools to collect and utilize a large speech data base for basic research in automatic speech recognition, under contract to Rome Air Development Center (1983 to 1986); FORTRAN-77 on VAX/VMS. Co-authored the final technical report on this research project.

  - Developed application software for Digital Video Interactive (DVI) technology, utilizing advanced interactive graphics, image processing, and motion video from compressed digital data (IR&D, 1988-89); "C" on PC/DOS.

ACADEMIC EXPERIENCE:

- Fall 1981 to Spring 1983: Research Fellow, engaged in Project on Urban Minorities and Language Change, directed by Dr. William Labov (Dept. of Linguistics). Responsibilities included: management/maintenance of computer resources; development of FORTRAN-IV software for sampling, playback and modified re- synthesis of speech; preparation of modified speech stimuli for human perception experiments; analysis and final report of experimental results.

- Spring 1984: Instructor for Quantitative Study of Sound Change (acoustic analysis of spontaneous speech, graduate level; Dept. of Linguistics).

- Fall 1983 to Fall 1995 (with some gaps): Dissertation research on acoustic variation in American English diphthongal vowels.

PAPERS PRESENTED OR PUBLISHED:

"The RATS Collection: Supporting HLT Research with Degraded Audio Data". David Graff, Kevin Walker, Stephanie Strassel, Xiaoyi Ma, Karen Jones, Ann Sawyer; LREC 2014: 9th Edition of the Language Resources and Evaluation Conference, Reykjavik, May 2014.

"Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement", Mohamed Maamouri, Wajdi Zaghouani, Violetta Cavalli-Sforza, David Graff, Mike Ciul; NAACL-HLT 2012: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, June 2012.

"New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus", Stephanie Strassel, Kevin Walker, Karen Jones, Dave Graff, Christopher Cieri; Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 2012.

"Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects", David Graff, Mohamed Maamouri; LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 2012.

"Speaker Recognition: Building the Mixer 4 and 5 Corpora", Linda Brandschain, Christopher Cieri, David Graff, Abby Neely, Kevin Walker; LREC 2008: 7th International Conference on Language Resources and Evaluation, Marrakech, May 2008.

"Lexicon Development for Varieties of Spoken Colloquial Arabic", by David Graff, Tim Buckwalter, Hubert Jin and Mohamed Maamouri; LREC 2006: Fifth International Conference on Language Resources and Evaluation, May 2006.

"Topic Detection and Tracking: Event-based Information Organization", by Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert and Mark Liberman; Kluwer International Series on Information Retrieval, Bruce Croft, ed. 2002.

"An overview of Broadcast News corpora", by David Graff; in Speech Communications, vol.37 pp.15-26; Elsevier, 2002.

"Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts", by Christopher Cieri, Dave Graff, Mark Liberman, Nii Martey and Stephanie Strassel; Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

"Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies", by David Graff and Steven Bird; Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

"Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora", Stephanie Strassel, Dave Graff, Nii Martey, Christopher Cieri; LREC 2000: 2nd International Language Resources and Evaluation Conference, Athens, May 2000.

"Public Data Bases for Speaker Recognition and Verification", by John Godfrey, David Graff and Alvin Martin; presented at the ESCA Workshop on Speaker Identification and Verification, April 1994.

"Testing listeners' reactions to phonological markers of ethnic identity", by D. Graff, W. Labov and W. A. Harris; Diversity and Diachrony, D. Sankoff, ed. 1986.