Kenneth W. Church

Email: kenneth.ward.church@gmail.com

Google Scholar (H-Index 70):
`https://scholar.google.com/citations?hl=en&user=E6aqGvYAAAAJ`

Education

PhD (1983)	in Computer Science	Massachusetts Institute of Technology
M.S. (1980)	in Computer Science	Massachusetts Institute of Technology
B.S. (1978)	in Computer Science	Massachusetts Institute of Technology

Employment

2022 - Present	Northeastern University, San Jose, CA
2018 - 2022	Baidu, Sunnyvale, CA
2011 - 2018	IBM TJ Wason, Yorktown Heights, NY
2009 - 2011	Johns Hopkins University, Baltimore, MD
2003 - 2009	Microsoft Research, Redmond, WA
1983 - 2003	AT&T Bell Labs, Murray Hill, NJ (and AT&T Labs, Florham Park)

Honors

2001	AT&T Fellow
1993 - 2011	President of ACL SIGDAT (organizes EMNLP)
2012	President of ACL
2015	ACL Fellow
2018	Baidu Fellow
2023	ACM Fellow

Advising Experience

Former students and post docs who went on to teaching positions:

- **Richard Sproat** (Google, but formally at UIUC, Linguistics): Post Doc at Bell Labs (1983).
- **Michel DeGraff** (MIT, Linguistics): summer intern (late 1980s).
- **David Yarowsky** (Johns Hopkins, Computer Science): one year at Bell Labs in early 1990s. Co-published approximately 10 papers between 1992 and 2011
- **Pascale Fung** (HKUST, EE): Summer Intern at Bell Labs (1993). Co-published papers in Coling-1994
- **Ido Dagan** (Bar Ilan, Computer Science): Post Doc at Bell Labs (mid 1990s). Co-published in ACL conferences in 1993 and 1994
- **Marti Hearst** (Berkeley, School of Information)
- **Ping Li** (CEO of a startup, but formally Rutgers, Statistics and CS): Intern at MSR (2004 & 2005)
  Co-published 7 papers including the *Best Student Paper* in KDD 2006.
- **Qiaozhu Mei** (Michigan, School of Information): Summer 2005, 2006 interns at MSR. Co-published papers in WSDM-2008 and CIKM-2008.

Representative Papers, selected from about 275 publications
Full list at `https://scholar.google.com/citations?hl=en&user=E6aqGvYAAAAJ`

1. I write 2-3 opinion pieces per year for the Journal of Natural Language Engineering. Usually 2 or 3 of these pieces are on the list of most read papers in the journal over the last 30 days: `https://www.cambridge.org/core/journals/natural-language-engineering/most-read`. The list of 20-some articles are: `https://www.cambridge.org/core/journals/natural-language-engineering/emerging-trends`.

2. I led a JSALT-2023 team in France on Deep Nets and Linear Algebra for applications in Academic Search. These are 6-week summer schools, that often produce highly-cited publications; see `https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html`. Mark Liberman and I led a JSALT team in 2017 on diarization which produced the DIHARD challenges; `https://arxiv.org/pdf/2012.01477.pdf` has 100 citations in Google Scholar.

3. **Kenneth Church** and Patrick Hanks, *Word association norms, mutual information, and lexicography*, Computational linguistics, 16:11, pp. 22–29, 1990, citations from Google Scholar: 6542. This paper introduced computational linguistics to what is now known as PMI (pointwise mutual information), which has direct connections to Word2Vec and Deep Nets such as BERT.

4. William Gale and **Kenneth Church**, *A method for aligning sentences in bilingual corpora,* Computational Linguistics, 19:1, pp. 75-102, 1993, citations from Google Scholar: 1878. Much of the recent progress in Machine Translation goes back to early work on parallel corpora.

5. **Kenneth Church**, *A stochastic parts program and noun phrase parser for unrestricted text*, Proceedings of the second conference on Applied natural language processing (an ACL conference), pp. 136–143, 1988, citations from Google Scholar: 1828. Part of Speech tagging was one of the early successes that led to a revival of empirical methods such as deep nets.

6. William Gale, **Kenneth Church** and David Yarowsky, *A method for disambiguating word senses in a large corpus,* Computers and the Humanities, 26:5–6, pp. 415–439, 1992, citations from Google Scholar: 917.

7. William Gale, **Kenneth Church** and David Yarowsky, *One sense per discourse,* Proceedings of the workshop on Speech and Natural Language, pp. 233–237, 1992, citations from Google Scholar: 809.

8. **Kenneth Church** and Robert Mercer, *Introduction to the special issue on computational linguistics using large corpora,* Computational linguistics, 19:1, pp. 1–24, 1993, citations from Google Scholar: 618.

9. **Kenneth Church**, *Emerging Trends: Word2Vec,* Natural Language Engineering, 23:1, 2017 citations from Google Scholar: 526.

10. Ping Li, Trevor Hastie and **Kenneth Church**, *Very sparse random projections*, KDD (best student paper), pp. 287–296, 2006, citations from Google Scholar: 487.

11. Qiaozhu Mei, Dengyong Zhou and **Kenneth Church**, *Query suggestion using hitting time,* CIKM, pp. 469–478, 2008, citations from Google Scholar: 774.

12. **Kenneth Church** and William Gale, *A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams,* Computer Speech & Language, 5:1, pp. 19–54, 1991, citations from Google Scholar: 370.

13. Mark Kernighan, **Kenneth Church** and William Gale, *A spelling correction program based on a noisy channel model* Coling, pp. 205–210, 1990, citations from Google Scholar: 414.

14. Mikio Yamamoto and **Kenneth Church**, *Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus,* Computational Linguistics, 27:1, pp. 1–30, 2001, citations from Google Scholar: 309.

15. **Kenneth Church** and Jonathan Helfman, *Dotplot: A program for exploring self-similarity in millions of lines of text and code,* Journal of Computational and Graphical Statistics, 2:2, pp. 153–174, 1993, citations from Google Scholar: 219.

16. **Kenneth Church**, Albert Greenberg and James Hamilton, *On Delivering Embarrassingly Distributed Cloud Services,* Hotnets, pp. 55–60, 2008, citations from Google Scholar: 206.

17. **Kenneth Church**, *Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p^2,* Coling, pp. 180–186, 2000, citations from Google Scholar: 213.

18. Aren Jansen, **Kenneth Church** and Hynek Hermansky, *Towards spoken term discovery at scale with zero resources,* INTERSPEECH, pp. 1676–1679, 2010, citations from Google Scholar: 181.

19. Adam Buchsbaum, Donald Caldwell, **Kenneth Church**, Glenn Fowler and S Muthukrishnan, *Engineering the compression of massive tables: an experimental approach,* SODA, 9:11, pp. 175–184, 2000, citations from Google Scholar: 80.

20. Qiaozhu Mei and **Kenneth Church**, *Entropy of search logs: how hard is search? with personalization? with backoff?* WSDM, pp. 45–54, 2008, video: `http://videolectures.net/wsdm08_mei_esl/` citations from Google Scholar: 83.

21. **Kenneth Church**, *A pendulum swung too far,* LiLT (Linguistic Issues in Language Technology), 2011, citations from Google Scholar: 98.

22. Ping Li and **Kenneth Church**, *A sketch algorithm for estimating two-way and multi-way associations,* Computational Linguistics, 33:3, pp. 305–354, 2007, citations from Google Scholar: 56.

23. **Kenneth Church**, *How many multiword expressions do people know?* TSLP (ACM Transactions on Speech and Language Processing), 10:2, 2013, citations from Google Scholar: 35.

Work Experience

- 2022-Present: Northeastern

  Joined a new group under Riccardo Baeza-Yates and Usama Fayyad, with interests in natural language, information retrieval and data mining.

- 2018-2022: Baidu

  Responsible for a small research team of machine learning experts with interests in language, vision and systems. Report to a senior vice president.

- 2011-2018: IBM

  Responsible for a small research team working on a small piece of Watson (customer care, medicine, drilling for oil). Analyzed logs for a few speech APIs.

- 2009-2011: Johns Hopkins University

  Chief Scientist of HLTCOE and Research Professor in Computer Science. Started work on zero resource speech recognition, showing that it is possible to do document retrieval on audio files in a surprise language with no resources (dictionaries and annotated corpora).

- 2003-2009: Microsoft Research

  I focused mostly on web search (e.g., publications with students including Ping Li and Qiaozhu Mei). But I also had the freedom to explore many other topics as well. My compression method was shipped in Microsoft Office for spelling correction (EMNLP-2007). I also had an opportunity to work with experts on cloud computing on embarrassingly distributed services (Hotnets-2008).

- 1983-2003: AT&T Bell Labs, Murray Hill (and AT&T Labs, Florham Park)

  I started as a member of technical staff and left as a fellow and a department head. Bell Labs encouraged interdisciplinary work. I published a number of papers with people in many fields: computer science (David Yarowsky), linguistics (Mark Liberman), statistics (Bill Gale) and lexicography (Patrick Hanks). I had the opportunity to work closely with the people who invented Unix.

  I spent much of the first decade at Bell Labs advocating the move in computatinal linguistics from rationalism (formalisms largely inspired by Chomsky) to empiricism (modern corpus-based methods). The second decade applied these methods to problems beyond speech and language (big data).

Important conference keynote talks (or invited talks)

- EACL-1993, SIGDAT-1999, AMTA-2002, Eurospeech-2003, LREC-2004, EMNLP-2004, TSD-2004 and TSD-2018 (Brno, Czech Republic), workshop in honor of Chuck Fillmore (ACL-2014), CCL-2019 (China), RANLP-2019 (Bulgaria), CLSW-2019 (China)

- Online Videos/Podcasts:

  `https://vimeo.com/37153276`, Hopkins CLSP Seminar (1999)
  `https://vimeo.com/37186100`, Hopkins CLSP Seminar (2003), rerun of keynote at Eurospeech-2003 (now known as Interspeech)
  `https://www.youtube.com/watch?v=EIVgGCSCCb4&t=2486s`, TSD Keynote (2018)
  `https://aneyeonai.libsyn.com/2019/05`, podcast (2019)
  `https://www.youtube.com/watch?v=lxwCymSbtJE`, Northeastern Seminar (2022)
  `https://github.com/kwchurch/ACL2022_deepnets_tutorial`, ACL Tutorial (2022)
  `https://www.youtube.com/@Kwchurch6340/videos` a channel with videos from workshops, etc.

Funding Experience

- I have been supported by industry for most of my career, except for a year at ISI (1990) and my time at Hopkins (2009-2011). At Hopkins, I was the Chief Scientist for HLTCOE (Human Language Technology Center of Excellence).

  The HLTCOE had a large grant from the Dept of Defence (approximately $9M per year for 9 years). During much of this time, while there was a search for a new director, I was responsible for the research program and the acting director was responsible for administrative functions.

- In addition, I was co-PI for a smaller $2.1M NSF grant. Data-Scope is a collaboration with astronomers and other scientists to make it easier to work with big data. See `https://magazine.krieger.jhu.edu/2011/10/data-scope-the-best-bar-none/`.

- I have reviewed for various NSF panels, as well as similar organizations in other countries. Now that I am back in academia, I have submitted a few proposals recently.