Kenneth W. Church

https://kwchurch.github.io/

Google Scholar (H-Index 72):
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, Boston, MA
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** (retired from 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** (Meta, but formally at 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 more than 300 publications
Full list at `https://scholar.google.com/citations?hl=en&user=E6aqGvYAAAAJ`

1. **Kenneth Church** and Patrick Hanks, *Word association norms, mutual information, and lexicography*, Computational linguistics, 16:11, pp. 22–29, 1990, citations from Google Scholar: 6834. This paper introduced computational linguistics to what is now known as PMI (pointwise mutual information), which has direct connections to Word2Vec and LLMs (large language models) such as BERT.

2. **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: 642. This paper came out at the time that we were starting EMNLP, now a major conference in Computational Linguistics. At that time the E-word (empiricism) was out-of-fashion. This paper (and EMNLP) helped change that. Mercer became an ACL Follow for his contributions to the use of machine learning in Speech Recognition and Machine Translation. Mercer has since become wealthy by using machine learning to trade stocks.

3. **Kenneth Church**, *A pendulum swung too far,* LiLT (Linguistic Issues in Language Technology), 2011, citations from Google Scholar: 109. This paper surveys the field from empiricism in 1950s (Shannon) to rationalism in 1970s (Minsky and Chomsky) to the revival of empiricism in 1990s (EMNLP) and predicts changes every 20 years. I predicted another big change in 2010s; as it turned out, that change would be LLMs.

4. **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: 1869. This was my first empirical paper. Before that, I was an advocate of rationalism following some of my teachers at MIT such as Minsky and Chomsky. Part of Speech tagging was one of the early successes that led to a revival of empirical methods such as LLMs.

5. **Kenneth Church**, *Emerging Trends: Word2Vec,* Natural Language Engineering, 23:1, 2017 citations from Google Scholar: 966. This is the most cited paper in my Emerging Trends column for the Journal of Natural Language Processing (formally the Journal of Natural Language Engineering). There are now 20-some articles,[1] mostly tutorials on Word2Vec, RAG and fine-tuning, as well as opinion pieces on benchmarking, reviewing, Responsible AI (ethics reviews, moderating of social media, proliferation of malware).

6. I led a JSALT-2023 team in France on Deep Nets and Linear Algebra for applications in Academic Search.[2] `http://recommendpapers.xyz` is a result of that effort. JSALT organizes 6-week summer schools every year for 3 decades, with an impressive track record of producing highly-cited publications, and identifying rising stars. Mark Liberman and I also led a JSALT team in 2017 on diarization which produced the DIHARD challenges; both the second and third challenges have about 200 citations in Google Scholar.

7. Ping Li, Trevor Hastie and **Kenneth Church**, *Very sparse random projections*, KDD (best student paper), pp. 287–296, 2006, citations from Google Scholar: 855. At the time, Ping Li was an intern working for me at Microsoft. He has been working on his intern project for the last 20 years, and is now the CEO of a start up company: `VecML.com`.

---

[1] `https://www.cambridge.org/core/journals/natural-language-engineering/emerging-trends`
[2] `https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html`

8. 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: 1927. Much of the recent progress in Machine Translation (MT) goes back to early work on parallel corpora. Bill Gale and I wrote a number of papers together on applications of statistics to speech and language including WSD (word sense disambiguation), Good-Turing Smoothing and spelling correction.

9. 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: 936. This was an early paper on WSD (word sense disambiguation) which helped create events such as SemEval.

10. 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: 438. At the time, Mark was in high-school. Jurafsky, who did not know this was a high-school project, included Mark's work in his textbook because it is a simple way to teach HMMs (Hidden Markov Models), an early version of deep nets.

11. Qiaozhu Mei, Dengyong Zhou and **Kenneth Church**, *Query suggestion using hitting time,* CIKM, pp. 469–478, 2008, citations from Google Scholar: 439. When I worked at Microsoft, I had access to search logs. A search engine is a market maker that tried to find an equilibrium between readers and writers, where crawls are closer to the writers' perspective(s), and the logs are closer to the readers' perspective(s). We talk about web crawls as large (and informative), but actually the logs are larger and more informative because there are more readers than writers (of successful content).

12. 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: 82.

13. 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. This is a very general approach for training LZ-compression (gzip) for tabular data (database records such as telephone call detail). Gzip is more effective if we enumerate tables to improve locality (similar content should appear close by). This method made it possible to ship all of AT&T's call detail across the country on a few T1 lines. Call detail is the data that is used to produce bills. Telephone call detail is similar to internet packet headers and logs for web search and social media. This kind of data can be used to learn more about users than they know about themselves.

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. Two of their sons were interns with me (Mark Kernighan and Peter McIlroy). I had many other interns as such as such Michel DeGraff and David Yarowksy; both are currently professors at MIT and Hopkins, respectively.

  I spent much of the first decade at Bell Labs advocating the move in computational 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). AT&T has some datasets that are much larger than natural language corpora such as call detail (meta data used to produce telephone bills).

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.[3]

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.

Coding Experience

Despite my day job, I continue to program. The academic search tool, `http:recommendpapers.xyz` is running on a small NAS in my house. A number of tutorial papers mentioned above are associated with repositories on `https://github.com/kwchurch`.

Conference/Workshop Organization

In addition to organizing many of the early EMNLP conferences (and earlier events known as the Workshop on Very Large Corpora), I have co-organized events such as a workshop on benchmarking at ACL-2021. Since that was during COVID, the talks and discussion were recorded and can be found on my GitHub.

---

[3]`https://magazine.krieger.jhu.edu/2011/10/data-scope-the-best-bar-none/`