Kenneth W. Church

https://kwchurch.github.io/

Contact

Home Page	https://kwchurch.github.io/
Email	Kenneth.Ward.Church@gmail.com
Mobile Phone	+1 4109496340
Address	80 E. Sunnyside Lane, apt #1, Irvington, NY, 10533, USA
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

Current Position and History

2025 - Present	VecML.com, Bellevue, WA, USA
2022 - 2025	Northeastern University, Boston, MA, USA
2018 - 2022	Baidu, Sunnyvale, CA, USA
2011 - 2018	IBM TJ Wason, Yorktown Heights, NY, USA
2009 - 2011	Johns Hopkins University, Baltimore, MD, USA
2003 - 2009	Microsoft Research, Redmond, WA, USA
1983 - 2003	AT&T Bell Labs, Murray Hill, NJ, USA (and AT&T Labs, Florham Park)

Personal Statement

I have been working for more than 45 years in Artificial Intelligence, Natural Language and Speech. For the last 35 years, I have been using **big data** to build large language models (**LLMs**). I have published more than 6 papers per year over that time, and double that more recently. I plan to keep that up for another decade. I currently work for VecML.com, a company founded by Ping Li, a former intern that has been working on his intern project for 20 years since winning the KDD-2006 best student paper award for Very sparse random projections. VecML does machine learning on vector databases. Chatbots such as ChatGPT and DeepSeek use vectors to represent text, speech, pictures, DNA **sequences**, etc. ANN (approximate nearest neighbor) methods answer questions of interest to many communities. While this technology is extremely exciting, I am concerned that **demand** may not be able to keep up with **supply**.

"Large" has always been a moving target. Back in the 1980s, we thought 250 million words of Associated Press Newswire was large. These days, Common Crawl, a popular dataset for training LLMs, is 45 TBs. The bottleneck used to be the cost of computing, but going forward, there will be concerns about peak data; companies like OpenAI are running out of data to train on. When I collaborated with the astronomer, Alex Szalay, at Hopkins, on Data-Scope, he was hoping my language data would be big and I was hoping his data would be astronomical. As it turns out, everything is growing exponentially, but few datasets are keeping up with improvements in computing (**1000x per decade**). His data are not limited by the size of the universe, but rather, by the size of budgets, whereas my data are limited by the population of the planet: there are only so many people and they have only so much time to talk. It is hard to find datasets that people care about that are growing faster than 1000x per decade. At Hopkins, which has a strong medical school, I learned about amazing improvements in DNA sequencing (**10,000x per decade**). I would love to

work with experts on sequencing data where demand (10,000x) is likely to out-pace supply (1000x) for the foreseeable future.

Five Most Relevant Publications

(out of more than 300 publications, including more than 40 patents)

1. Ping Li, Trevor Hastie and **Kenneth Church**, Very sparse random projections, KDD (best student paper), pp. 287–296, 2006, citations from Google Scholar: 894. Ping Li has been working on his intern project for 20 years. In 2006, Ping Li was working for me as an intern at Microsoft Research, though I now work for him at his startup company: VecML.com.

2. **Kenneth Church** and Patrick Hanks, Word association norms, mutual information, and lexicography, *Computational linguistics*, 16:11, pp. 22–29, 1990, citations from Google Scholar: 6955. 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 and **bots** such as ChatGPT and DeepSeek.

3. **Kenneth Church**, From PMI to Bots, *International Journal of Lexicography*, 2025. This recent paper, published in a special issue in memory of Patrick Hanks, connects the dots between PMI and much of the recent excitement in Artificial Intelligence (e.g., Word2Vec, LLMs, ChatGPT, DeepSeek).

4. **Kenneth Church**, Emerging Trends: Word2Vec, *Natural Language Engineering,* 23:1, 2017 citations from Google Scholar: 1140. 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, mostly tutorials on Word2Vec, RAG and fine-tuning, as well as opinion pieces on benchmarking, reviewing and Responsible AI (ethics, moderating of social media, proliferation of malware).

5. **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: 659. 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. In 1988, there were almost no statistical papers in ACL. A decade later, there were almost no non-statistical papers. Mercer won the ACL Lifetime Achievement Award for his contributions to the use of machine learning in Speech Recognition and Machine Translation. Mercer has since become wealthy by using **machine learning** and **big data** to trade stocks. It has been reported in the Guardian and elsewhere that Mercer used similar methods to influence elections (Brexit and Trump). The methods discussed in our paper are very powerful and can be used to do many things that go well beyond natural language (for better and for worse).

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

I have a number of former interns and post docs with impressive publication records. My two most recent post docs, Ibrahim Said Ahmad and John Ortega, are minorities. They are both interested in low resource languages. Ibrahim is a native speaker of Hausa, a language that I prefer to refer to as a growth opportunity because of the large number of speakers of Hausa as well as the impressive GDP growth in Nigeria. John is interested in a Quechua, a widely-spoken Indigenous language in Peru. One of my first interns, Michel DeGraff, a native speaker of Haitian Creole, is currently a professor at MIT. I have also had the honor to mentor a number of women who are now important leaders in computational linguistics such as Pascale Fung and Marti Hearst. Both of their h-indexes are higher than mine. A few other (non-minority) former post docs and interns are: David Yarowsky, Richard Sproat, Ido Dagan, Qiaozhu Mei and Ping Li. All of them are currently professors, or have been professors.

Leadership

I have held a number of management positions at AT&T Bell Labs, Hopkins and Baidu.

1. At AT&T Bell Labs, I was a department head.
2. At Hopkins, I was the Chief Scientist of the HLT COE (Human Language Technology Center of Excellence). 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.
3. At Baidu, I reported to a Senior Vice President and attended the quarterly directors' meeting in China. Among other things, I organized the advisory board for Baidu Research.

More recently, I led a JSALT-2023 team in France on Deep Nets and Linear Algebra for applications in Academic Search. As a result of that effort, there is a website for recommending papers, `http://recommendpapers.xyz`, running on an inexpensive NAS box in my house. Disk space is much cheaper at the edge (in my house) than in the cloud. This website uses considerable disk space since it is indexing more than 200 million academic papers from Semantic Scholar. 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; the first, second and third challenges have 100 or more citations in Google Scholar.

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.

Teaching

I taught CS 7290 – Special Topics in IR and NLP three times with Omar Alonso. We focus on presentation skills including a variety of formats, purposes and audiences:

1. formats: oral, written, video, poster, demos
2. purposes: proposals, status updates, final reports, interviews
3. audiences: technical conferences, funding agencies, VCs, industrial R&D

I taught a similar class at Columbia when I was working for IBM in New York.

I have also taught CS6120: Practical Natural Language Processing at Northeastern. This is a graduate class on natural language and LLMs.

At Hopkins, I taught an introductory undergraduate class on computer science for non-majors. Just as math departments teach calculus to a broad audience of non-majors, there should be a similar class that teaches computational thinking to non-majors. I taught this class a couple of times, once with Ann Irvine; it can be a challenge to increase female enrollment in CS, but she was super-helpful.

Soon after Mitch Marcus moved from Bell Labs to Penn, I taught a class at Penn on empirical methods. The class included a number of professors, as well as graduate students such as Michael Collins that were about to play a leading role in moving the field from rationalism to empiricism.

I have given many tutorials at conferences such as ACL, CIKM, WSDM and summer schools. Unix for Poets was developed for a linguistics summer school more than 20 years ago to convince them to write programs. It is amazing that these slides can still be found in relatively recent blogs such as this. Unix pipes and regular expressions work surprisingly well with linguists, though when teaching programming to freshman at Hopkins (who plan to study medicine), they tend to find Python more intuitive.

Funding

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) and Northeastern (2022-2025). At Hopkins, I was the Chief Scientist for HLTCOE (Human Language Technology Center of Excellence). The HLTCOE had a large grant from the Department of Defense (approximately $9M per year for 9 years).

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.

I have reviewed for various NSF panels, as well as similar organizations in other countries. Now that I am back in academia, I am working on a number of proposals including:

1. AI Meets Pharma: Autonomous Agents for Rapid, End-to-End Drug Development,
2. porting LLMs to growth languages (widely spoken languages with impressive GDP growth and relatively few resources), and
3. using nuclear norms to evaluate applications of LLMs to different samples of inputs.

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 (with 68 TBs of disk space). Much of the code behind that website is available from https://github.com/kwchurch. That GitHub also contains code for a number of tutorials mentioned above. A number of tutorial papers mentioned above are associated.