jeremy@trochee.net
@trochee
http://trochee.net/about

4549 48th Ave. SW
Seattle, WA 98116
cell: 206-650-3615

Jeremy G. Kahn, Ph.D.

Empiricist. Software engineer. Computational linguist.

Specialization

purpose Finding & answering interesting questions in rich, complex, noisy data, with special attention to natural language and human interaction.

instrument Scientific understanding, playfulness, curiosity, and software tooling to support all of the above.

theme Building and applying maintainable, understandable, extensible scientific and exploratory software that empowers and delights.

manner Marrying engineering-, science-, and humanity-based ideas of empiricism, information, utility, and explanatory power.

Employment

2014–2016

- **Senior Software Engineer**, Google, Seattle, Washington, in ~~(group name redacted)~~, within Research & Machine Intelligence. Prototyping and development of distributed (and non-distributed) neural networks, most concretely for language modeling. Worked on challenges of privacy, client-side machine intelligence, and distributed modeling and personalization.

  - ETL from real-world-scale text corpora (mostly in C++)
  - Recurrent network design and parametrization (mostly in Python)
  - Tools and abstractions for network training and evaluation of hyperparameters (mostly in Python and Jupyter(-like) notebooks).

  Collaborated with sibling teams working on ~~(text-focused products)~~ and image processing to explore the utility of client-side MI. Various projects to integrate with mobile work in Android-compatible Java.

Employment (continued)

2012–2013 • **Software Architect**, inome Inc. & **Senior Scientist, Data**, Intelius, Inc.; Bellevue, Washington. Ambidextrous in scientific and engineering work. Main contributions in both departments:

- Improved & refactored legacy Python/Java hybrid machine-learning system into largely independent, testable domain-knowledge and machine-learning modules.
- Introduced API-first and artifact-driven development, advocated and implemented in favor of a move from a monolithic build/deploy Hadoop-only practice to a `git`-federated (API-first, and artifact-driven) shared-best-practice model using Avro protocols and a service-oriented architecture.
- Revised core model of data store from a flat record structure with one datatype to a graph-database-backed approach (using Tinkerpop APIs). Led efforts to port, refactor and replace legacy components in the transition to new data model and interaction mechanics.
- Introduced & advocated for robust software and algorithm development strategies to both engineering and research teams. Design and implementation work on protocols, release mechanics, testing strategy, performance and quality metrics, and algorithm/architecture tensions.

2011–2012 • **Senior Computational Linguist**, Natural Language Processing, Quid Inc., San Francisco, California. Development and design of natural language processing systems and algorithms, including information extraction and processing systems. Software design and implementation. Research direction and systems planning. Git and Python hacker.

2011 • **Research Linguist**, Speech Technology and Research (STAR) Laboratory, SRI International, Menlo Park, California. Supporting two projects: GALE identification of non-standard Arabic dialect and ALADDIN identification of acoustic events in video. Doing work in text processing, data analysis, systems design and evaluation. Perl hacking of complex legacy systems.

2009–2011 • **Computational Linguist & Director of Research**, Wordnik, San Mateo, California. Applied statistical and knowledge-based linguistics to the research and development division of this computational lexicography startup. Planning and design work for advanced computational-linguistic features for the site and API.
Research and development on distributional semantics, multi-word expression, NLP systems design and implementation, machine learning, and more.

2008–2010 • **Visiting Fellow**, Speech Technology and Research (STAR) Laboratory, SRI International, Menlo Park, California. Researcher in support of the DARPA-funded GALE project.
Issues: Continuing research on using parsers in support of speech recognition, machine translation, and machine translation evaluation. Text-reconciliation for Mandarin Chinese among the various components of our GALE team, and the unofficial on-site representative for SSLI at SRI.

Employment (continued)

2003–2010

- **Research Assistant**, Signal, Speech and Language Interpretation (SSLI) Laboratory, Department of Electrical Engineering, University of Washington, Seattle. Supervisor: Dr. Mari Ostendorf.
  Machine-learning algorithms for the automatic processing of speech and text in English and other languages (especially Mandarin Chinese). Statistical and rule-based methods, with an emphasis on data-driven approaches, scientific engineering, and the incorporation of information theory and linguistic knowledge into these approaches. Informally, spokesman for software engineering and cluster-use best-practices and linguistics exposition.

2000–2003

- **Computational Linguist and Software Engineer**, Conversay Corporation, Redmond, Washington. Built multi-language, portable, re-usable text-to-speech and speech-recognition software.
  Issues: text-processing, NLP pipeline design, pronunciation prediction, data management, character-encoding concerns, software architecture and mutual evangelism among linguists, speech technologists, and software engineers.

1998–2000

- **Linguist/Programmer**, Eloquent Technology International, Ithaca, New York.
  Issues: pronunciation prediction, part-of-speech tagging, testing, software architecture, source-control, and meta-language optimization.

Education

2005–2010

- **Ph.D., Computational Linguistics**. University of Washington, Seattle.
  Dissertation: “Parse decoration of the word sequence in the speech-to-text machine-translation pipeline”. Using statistical parsers of English and Mandarin Chinese to:

  - improve large-vocabulary speech recognition results,
  - improve (statistical) machine translation word-alignment of those results, and
  - automatically evaluate translation candidates.

  Advisor: Professor Mari Ostendorf, (Electrical Engineering, adjunct in Linguistics and in Computer Science and Engineering).

- Coursework in applications of statistical methods and machine-learning to natural language processing (NLP) and speech applications.

2003–2005

- **M.A., Linguistics**. University of Washington, Seattle.
  Thesis: “Moving beyond the lexical layer in parsing conversational speech”. Improving statistical parsers over automatically-transcribed speech by:

  - using non-lexical speech cues like intonational and sentence boundaries, and
  - considering alternate word hypotheses when parsing

- Coursework in formal linguistics, machine learning, grammar engineering, statistical language processing.
- Founding (and current) member, UW Computational Linguistics Laboratory.
- Awarded Departmental Research Assistantship (declined).

Education (continued)

2002
- **Certificate, Foundations of Computer Systems**. Classes in discrete math, data structures and algorithms.

1993–1997
- **A.B. *magna cum laude*, Linguistics**. Brown University, Providence, Rhode Island.

Selected Publications

2010
- Jeremy G. Kahn and Mari Ostendorf (2010). Joint reranking of parsing and word recognition with automatic segmentation. *Computer Speech and Language.*

2009
- Jeremy G. Kahn, Matthew Snover, and Mari Ostendorf (2009). Expected dependency pair match: Predicting translation quality with expected syntactic structure. *Machine Translation.*

2006
- William P. McNeill, Jeremy G. Kahn, Dustin, L. Hillard, and Mari Ostendorf (2006). Parse structure and segmentation for improving speech recognition. In *Proceedings of the Conference on Speech and Language Technologies.*

Selected open source contributions

- **Packager, maintainer.** Pronunciation dictionary packaged for Python tools.

Samyro
- **Author.** RNN-based text generation engine and tools based on the TensorFlow C++/Python libraries.

EDPM
- **Author.** The Expected Dependency-Pair Match machine-translation evaluation library from the 2008 MetricsMATR competition.

Avro (Apache)
- Data serialization and protocol definition system. Contributions focused on Python interoperability and speed.

Pydoop
- Python/C libraries for Python interaction with Hadoop Pipes. Contributions support use of Pydoop with multiple input directories (for “reduce join” application).

`Hadoop`
- Python library for interacting with Java-native sequence files. Generalized file-read functionality to work with Pydoop.

Group Memberships

2000–present
- Association for Computational Linguistics (ACL), member.

2001–present
- Comprehensive Perl Archive Network (CPAN), contributor.

2003–present
- Linguistics Society of America, member.

2013–present
- Pypi package maintainer.

2014–present
- Association for Computing Machinery (ACM), member.