Collin Brown

Resources on Computational Linguistics

This is simply going to be a repository where I throw links and whatnot that I find in my perusing—anything that I want to hold on to for later that has any connection to computational linguistics. If you think there's something I should add to this list, feel free to email me.

I'm currently preparing to apply to master's programs in computational linguistics, and I thought it might be useful to peruse the internet and see what I couldn't dig up on the topic. I'm especially interested in the application of machine learning to historical linguistics—a topic I'm writing an article on at the moment, in fact—so it may be the case that many of my resources will focus on that. However, I'll try to keep my net as broad as possible. In any case, I hope this list proves useful.

I recently reached out to Nathan Schneider, an assistant professor of linguistics and computer science at Georgetown University, and he advised I read Jurafsky and Martin's *Speech and Language Processing*. The whole thing is available online, though it doesn't appear to be entirely finished. Jurafsky and another Stanford professor, Chris Manning, have a collection of a hundred or so videos on natural language processing that cover all the essentials.

Dr. Schneider also curates a list of linguistics departments that do computational research, particularly natural language processing—you can find it on his site or on github.

The Natural Language Toolkit is a “free, open source, community-drive” platform for “building Python programs to work with human language Data.” It has a nice guide that’ll teach you Python essentials and is all-around quite approachable.

Another book, this time on the subject of deep learning, is the aptly titled, *Deep Learning*, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville—a fairly general look at the field.

There are also a handful of repositories such as Awesome-NLP and Awesome for Deep Learning NLP that aggregate information on NLP and deep learning-related topics. For libraries and whatnot, I might point you towards LazyNLP, a site scraper that helps you accumulate large language datasets; Ciphey, a decryption tool that makes use of natural language processing; Doccano, an open-source annotation tool that helps build datasets for nlp tasks; and TextBlob, a textual data-processing library.

If you’re particularly interested in Basque—and why would you not be—you might take a look at the Ixa Group site, associated with the University of the Basque Country, which aggregates news related to Basque computational linguistics and the like.

If you’re more of a blog-reader, then these might be your cup of tea:

- Université du Québec's Daniel Lemire has a résumé that'd humble God himself and a nice blog to boot. He seems more focused on general computer science and software engineering, but that sort of thing can prove quite useful for our purposes. He also has a list of recommended video games and a guide on how to write well—a man after my own heart.
- The nlp-focused blog of Hal Daumé III—a senior principal researcher at Microsoft, professor at the University of Maryland, and assistant professor at the University of Utah.
- The much-acclaimed blog of Sebastian Ruder, a research scientist at Google; I might also point you to his repository of natural language processing improvements which should indicate what constitutes the "state-of-the-art" right now. Ruder recommends the Goodfellow, Bengio, and Courville book, *Deep Learning*, if you're interested in entering into the field. He also advises doing Steven Ng's Coursera course on machine learning and points the reader towards fast.ai, which seems to have a good course particularly on deep learning.
- The blog of Dirk Hovy, an associate professor at Bocconi University, who focuses on natural language processing. He has a repository, Python for Linguists - A Gentle Introduction to Programming , that might prove useful if you're coming from a linguistics background, and he provides access to a whole lot of his papers directly on his site.

- Andrej Karpathy—who helped found OpenAI and was the senior director of AI at Tesla—has a blog where he talks quite a lot about machine learning, in particular deep learning.
- The blog of Deniz Yuret, a professor at Koç University in İstanbul—I ran into his work ultimately through a paper he published, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." He does have some Turkish language resources which might be useful if that's something you're interested in.
- Vered Shwartz is an assistant professor at the University of Columbia; she seems to have a number of posts on natural language processing that might prove interesting.
- The LingPipe Blog focuses on "natural language processing and text analytics" which seems right up our alley. They also provide access to quite a few papers on various subjects that might prove useful or at least entertaining.
- John Langford, a partner researcher manager at Microsoft, seems like a member of those now rare breeds of scienticians whose sites feel homely in a strange, sort of frontiering way. His old site is even better. But I'm actually more interested in hunch.net, an experimental blog he started on machine learning. A few notable names have been published there, though most the links to their individual sites have rotted. I was able to find links to the works of Sanjoy Dasgupta and to the site of Alexander Gray. It seems our old friend, Hal Daumé III, was involved as well.

Now we're moving into more adjacent fields. I'm going to include some plain old linguistics and computer science blogs that might not be directly applicable to the subject but may still prove interesting.

- Christopher Olah's blog on neural networks; he has worked at OpenAI and Google, and while his posts seem to focus on subjects adjacent to our interest, it would do one well to learn about such things so I figured he was worth mentioning here.
- If you're interested in something a bit more business-focused, you can head over to the blog of Kavita Ganesan, a self-described AI advisor who has worked at Opinosis Analytics, eBay, Github, and the Huntsman Cancer Institute. Her blog seems to focus on general AI and their incorporation into businesses, so again not particularly applicable to what we're looking at here but perhaps worth mentioning.
- Ben Frederickson's seems to be more of a general software blog, but being a sucker for this sort of thing I can't help but include it here.
- I can't help but include the blog of Logan Kearsley, a fellow member of the Language Creation Society, who to my knowledge focuses more on constructed languages in his posts but may occasionally touch on more computational matters.

This is obviously an ongoing list that I'll be adding to here and there as I accumulate more resources. In any case, thanks for reading this far. Hopefully it has proved useful or at least a little entertaining. Obviously, I've only aggregated a hair-thin sliver of all the resources that are surely out there, but you have to start somewhere.