The Case for Empiricism (With and Without Statistics)

Kenneth Church
1101 Kitchawan Road
Yorktown Heights, NY 10589
USA
Kenneth.Ward.Church@gmail.com

Abstract

These days we tend to use terms like *empirical* and *statistical* as if they are interchangeable, but it wasn't always this way, and probably for good reason. In *A Pendulum Swung Too Far* (Church, 2011), I argued that graduate programs should make room for both Empiricism and Rationalism. We don't know which trends will dominate the field tomorrow, but it is a good bet that it won't be what's hot today. We should prepare the next generation of students for all possible futures, or at least all probable futures. This paper argues for a diverse interpretation of Empiricism, one that makes room for everything from Humanities to Engineering (and then some).

Figure 1: Lily Wong Fillmore (standing) and Charles (Chuck) Fillmore

1 Lifetime Achievement Award (LTA)

Since the purpose of this workshop is to celebrate Charles (Chuck) Fillmore, I would like to take this opportunity to summarize some of the points that I made in my introduction to Chuck's LTA talk at ACL-2012.

I had the rather unusual opportunity to see his talk (a few times) before writing my introduction because Chuck video-taped his talk in advance.[1] I knew that he was unable to make the trip, but I had not appreciated just how serious the situation was. I found out well after the fact that the LTA meant a lot to him, so much so that he postponed an operation that he probably shouldn't have postponed (over his doctor's objection), so that he would be able to answer live questions via Skype after the showing of his video tape.

I started my introduction by crediting Lily Wong Fillmore, who understood just how much Chuck wanted to be with us in Korea, but also, just how impossible that was. Let me take this opportunity to thank her once again for her contributions to the video (technical lighting, editing, encouragement and so much more).

For many of us in my generation, C4C, Chuck's "The Case for Case" (Fillmore, 1968) was the introduction to a world beyond Rationalism and Chomsky. This was especially the case for me, since I was studying at MIT, where we learned many things (but not Empiricism).

After watching Chuck's video remarks, I was struck by just how nice he was. He had nice things to say about everyone from Noam Chomsky to Roger Schank. But I was also struck by just how difficult it was for Chuck to explain how important C4C was (or even what it said and why it mattered). To make sure that the international audience wasn't misled by his upbringing and his self-deprecating humor, I showed a page of "Minnesota Nice" stereotypes, while reminding the audience that stereotypes aren't nice, but as stereotypes go, these stereotypes are about as nice as they get.

[1] The video is available online at https://framenet.icsi.berkeley.edu/fndrupal/node/5489.

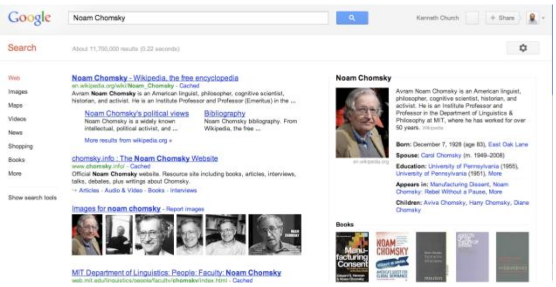
Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), pages 6–9,
Baltimore, Maryland USA, June 27, 2014. ©2014 Association for Computational Linguistics

Chuck, of course, is too nice to mention that Fillmore (1967) had 6000 citations in Google Scholar as of ACL-2012.[2] He also didn't mention that he has another half dozen papers with 1000 or more citations including an ACL paper on FrameNet (Baker *et al*, 1998).[3]

I encouraged the audience to read C4C. Not only is it an example of a great linguistic argument, but it also demonstrates a strong command of the classic literature as well as linguistic facts. Our field is too "silo"-ed. We tend to cite recent papers by our friends, with too little discussion of seminal papers, fields beyond our own, and other types of evidence that go beyond the usual suspects. We could use more "Minnesota Nice."

I then spent a few slides trying to connect the dots between Chuck's work and practical engineering apps, suggesting a connection between morphology and Message Understanding Conference (MUC)-like tasks. We tend to think too much about parsing (question 1), though question 2 is more important for tasks such as information extraction and semantic role labeling.

1. What is the NP (and the VP) under S?
2. Who did what to whom?



Noam Chomsky

About 11,700,000 results (0.22 seconds)

Noam Chomsky - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Noam_Chomsky - Cached
Avram **Noam Chomsky** is an American linguist, philosopher, cognitive scientist, historian, and activist. He is an Institute Professor and Professor (Emeritus) in the ...

Noam Chomsky's political views
Noam Chomsky is a widely known intellectual, political activist, and ...

Bibliography
Noam Chomsky bibliography. From Wikipedia, the free ...

More results from wikipedia.org »

chomsky.info : The **Noam Chomsky** Website
www.chomsky.info/ - Cached
Official **Noam Chomsky** website. Resource site including books, articles, interviews, talks, debates, plus writings about Chomsky.
→ Articles - Audio & Video - Books - Interviews

Images for **noam chomsky** - Report images

MIT Department of Linguistics: People: Faculty: **Noam Chomsky**
web.mit.edu/linguistics/people/faculty/**chomsky**/index.html - Cached

Noam Chomsky

Avram Noam Chomsky is an American linguist, philosopher, cognitive scientist, historian, and activist. He is an Institute Professor and Professor in the Department of Linguistics & Philosophy at MIT, where he has worked for over 50 years. Wikipedia

en.wikipedia.org

Born: December 7, 1928 (age 83), East Oak Lane

Spouse: Carol Chomsky (m. 1949-2008)

Education: University of Pennsylvania (1955), University of Pennsylvania (1951), More

Appears in: Manufacturing Dissent, Noam Chomsky: Rebel Without a Pause, More

Children: Aviva Chomsky, Harry Chomsky, Diane Chomsky

Books

Figure 2: An example of information extraction in commercial practice.

Context-Free Grammars are attractive for languages with more word order and less morphology (such as English), but Case Grammar may be more appropriate for languages with more morphology and less word order (such as Latin, Greek & Japanese). I then gave a short (over-simplified) tutorial on Latin and Japanese grammar, suggesting a connection between Latin cases (e.g., nominative, accusative, ablative, etc.) and Japanese function words (e.g., the subject marker *ga* and the direct object marker *wo*, etc.). From there, I mentioned a few historical connections

- Case Grammar → Frames → FrameNet
- Valency[4] → Scripts (Roger Schank)
- Chuck → Sue Atkins (Lexicography)

The verb "give," for example, requires three arguments: *Jones* (agent) *gave money* (object) *to the school* (beneficiary). In Latin, these arguments are associated with different cases (nominative, accusative, etc.). Under the frame view, similar facts are captured with a *commercial transaction frame*, which connects arguments across verbs such as: *buy, sell, cost* and *spend*.[5]

VERB	**BUYER**	**GOODS**	**SELLER**	**MONEY**	**PLACE**
buy	subject	object	from	for	at
sell	to				
cost	indirect object	subject		object	at
spend	subject	on		object	at

Lexicographers such as Sue Atkins use patterns such as:

- Risk <valued object> for <situation> | <purpose> | <beneficiary> | <motivation>

to address similar alternations. My colleague Patrick Hanks uses a similar pattern to motivate our work on using statistics to find collocations:

- Save <good thing> from <bad situation>

Lexicographers use patterns like this to account for examples such as:

- *Save whales from extinction*
- *Ready to risk everything for what he believes.*

where we can't swap the arguments:

- **Save extinction from whales*

The challenge for the next generation is to move this discussion from lexicography and general linguistics to computational linguistics. Which of these representations are most appropriate for practical NLP apps? Should we focus on part of speech tagging statistics, word order or frames

[2] Citations tend to increase over time, especially for important papers like Fillmore (1967), which has more than 7000 citations as of April 2014.

[3] See framenet.icsi.berkeley.edu for more recent publications such as Ruppenhofer *et al* (2006).

[4] http://en.wikipedia.org/wiki/Valency_(linguistics)

[5] For more discussion of this table, see www.uni-stuttgart.de/ linguistik/ sfb732/ files/ hamm_framesemantics.pdf

(typical predicate-argument relations and collocations)?

Do corpus-based lexicography methods scale up? Are they too manually intensive? If so, could we use machine learning methods to speed up manual methods? Just as statistical parsers learn phrase structure rules such as S → NP VP, we may soon expect machine learning systems to learn valency, collocations and typical predicate-argument relations.

How large do the corpora have to be to learn what? When can we expect to learn frames? In the 1980s, corpora were about 1 million words (Brown Corpus). That was large enough to make a list of common content words, and to train part of speech taggers. A decade later, we had 100 million word corpora such as the British National Corpus. This was large enough to see associations between common predicates and function words such as "save" + "from." Since then, with the web, data has become more and more available. Corpus growth may well be indexed to the price of disks (improving about 1000x per decade). Coming soon, we can expect $1M^2$ word corpora. (Google may already be there.) That should be large enough to see associations of pairs of content words (collocations). At that point, machine learning methods should be able to learn many of the patterns that lexicographers have been talking about such as: *risk valued object for purpose*.

We should train the next generation with the technical engineering skills so they will be able to take advantage of the opportunities, but more importantly, we should encourage the next generation to read the seminal papers in a broad range of disciplines so the next generation will know about lots of interesting linguistic patterns that will, hopefully, show up in the output of their machine learning systems.

2 Empirical / Corpus-Based Traditions

As mentioned above, there is a direct connection between Fillmore and Corpus-Based Lexicographers such as Sue Atkins (Fillmore and Atkins, 1992). Corpus-based work has a long tradition in lexicography, linguistics, psychology and computer science, much of which is documented in the Newsletter of the International Computer Archive of Modern English (ICAME).[6] According to Wikipedia,[7] ICAME was co-founded by Nelson Francis, who is perhaps best known for his collaboration with Henry Kučera on the Brown Corpus.[8] The Brown Corpus dates back to the 1960s, though the standard reference was published two decades later (Francis and Kučera, 1982).

The Brown Corpus has been extremely influential across a wide range of fields. According to Google Scholar, the Brown Corpus has more than 3000 citations. Many of these references have been extremely influential themselves in a number of different fields. At least[9] ten of these references have at least 2000 citations in at least five fields:

- Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999),
- Lexicography (Miller, 1995),
- Sociolinguistics (Biber, 1991),
- Psychology (MacWhinney, 2000)
- Computational Linguistics (Marcus *et al*, 1993; Jurafsky and Martin, 2000; Church and Hanks, 1990; Resnik, 1995)

All of this work is empirical, though much of it is not all that statistical. The Brown Corpus and corpus-based methods have been particularly influential in the Humanities, but less so in other fields such as Machine Learning and Statistics. I remember giving talks at top engineering universities and being surprised, when reporting experiments based on the Brown Corpus, that it was still necessary in the late 1990s to explain what the Brown Corpus was, as well as the research direction that it represented. While many of these top universities were beginning to warm up to statistical methods and machine learning, there has always been less awareness of empiricism and less sympathy for the research direction. These days, I fear that the situation has not improved all that much. In fact, there may be even less room than ever for empirical work (unless it is statistical).

It is ironic how much the field has changed (and how little it has changed). Back in the early 1990s, it was difficult to publish papers that digressed from the strict rationalist tradition that dominated the field at the time. We created the Workshop on Very Large Corpora (WVLC

[6] http://icame.uib.no/archives/No_1_ICAME_News.pdf

[7] http://en.wikipedia.org/wiki/W._Nelson_Francis

[8] http://en.wikipedia.org/wiki/Brown_Corpus

[9] Google Scholar is an amazing resource, but not perfect. There is at least one error of omission: Manning and Schütze (1999).

evolved into EMNLP) to make room for a little work of a different kind. But over the years, the differences between the main ACL conference and EMNLP have largely disappeared, and the similarities between EMNLP and ICAME have also largely disappeared. While it is nice to see the field come together as it has, it is a shame that these days, it is still difficult to publish a paper that digresses from the strict norms that dominate the field today, just as it used to be difficult years ago to publish papers that digressed from the strict norms that dominated the field at the time. Ironically, the names of our meetings no longer make much sense. There is less discussion than there used to be of the E-word in EMNLP and the C-word in WVLC.

One of the more bitter sweet moments at a WVLC/EMNLP meeting was the invited talk by Kučera and Francis at WVLC-1995, [10] which happened to be held at MIT. Just a few years earlier, it would have been unimaginable that such a talk could have been so appreciated at MIT of all places, given so many years of such hostility to all things empirical.

Their talk was the first and last time that I remember a standing ovation at WVLC/EMNLP, mostly because of their contributions to the field, but also because they both stood up for the hour during their talk, even though they were well past retirement (and standing wasn't easy, especially for Francis).

Unfortunately, while there was widespread appreciation for their accomplishments, it was difficult for them to appreciate what we were doing. I couldn't help but notice that Henry was trying his best to read other papers in the WVLC-1995 program (including one of mine), but they didn't make much sense to him. It was already clear then that the field had taken a hard turn away from the Humanities (and C4C and FrameNet) toward where we are today (more Statistical than Empirical).

3 Conclusion

Fads come and fads go, but seminal papers such as "Case for Case" are here to stay. As mentioned above, we should train the next generation with the technical engineering skills to take advantage of the opportunities, but more importantly, we should encourage the next generation to read seminal papers in a broad range of disciplines so they know about lots of interesting linguistic patterns that will, hopefully, show up in the output of their machine learning systems.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto.1999. *Modern information retrieval*. Vol. 463. ACM Press, New York, NY, USA.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. "The berkeley framenet project," *ACL.*

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Kenneth Church. 2011. A pendulum swung too far, *Linguistic Issues in Language Technology*, 6(5).

Kenneth Church and Patrick Hanks. 1990 "Word association norms, mutual information, and lexicography." *Computational linguistics* 16(1): 22-29

Charles J. Fillmore. 1968. "The Case for Case." In Bach and Harms (Ed.): *Universals in Linguistic Theory.* Holt, Rinehart, and Winston, New York, NY, USA, pp. 1-88.

Charles J. Fillmore and Beryl TS Atkins. 1992. "Toward a frame-based lexicon: The semantics of RISK and its neighbors." *Frames, fields, and contrasts*, pp. 75-102, Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

W. Nelson Francis and Henry Kučera. 1982 *Frequency analysis of English usage.* Houghton Mifflin, Boston, MA, USA.

Dan Jurafsky and James H. Martin. 2000 *Speech & Language Processing*. Pearson Education India.

Brian MacWhinney. 2000. *The CHILDES Project: The database*. Vol. 2. Psychology Press.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press. Cambridge, MA, USA.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. "Building a large annotated corpus of English: The Penn Treebank." *Computational linguistics* 19(2): 313-330.

George A. Miller. 1995. "WordNet: a lexical database for English." *Communications of the ACM* 38(11): 39-41.

Philip Resnik. 1995. "Using information content to evaluate semantic similarity in a taxonomy." *arXiv preprint cmp-lg/9511007*

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended theory and practice*. framenet.icsi.berkeley.edu

[10] http://aclweb.org/anthology//W/W95/W95-0100.pdf