METHODS

Dialogic Density in the Books of William James and John Dewey

Alexander Klein, McMaster University

Study Overview

I employed a machine learning algorithm (the Stanford Named Entity Recognizer, or NER; see Finkel, Grenager, and Manning 2005) to examine the relative rates at which William James and John Dewey mention other persons in representative samples of their respective writings. The NER attempts to tag words and phrases in a corpus with either PERSON, ORGANIZATION, or LOCATION. I created two corpora, one for each author's key monographs and mono-authored collections. I then assembled databases for each book in each corpus. The databases collect every PERSON tag found using NER in each book. This allows me to model the relationship between the rate at which names of persons appear in each corpus per every 1000 words. I call this rate a corpus's "dialogic density."

Why model this relationship? In philosophy, pragmatists like James and Dewey both have a reputation as radicals who are more interested in changing the subject on old philosophical problems than in engaging. For experienced readers of James, however, this description might seem less apt, since his work is saturated in references to other persons, including to their ideas, theories, and data. This study aims to assess whether there are significant differences in the rates at which these two philosophers mention other figures in their writing. My hypothesis is that James is significantly more likely to reference other persons in his writing than is Dewey.

I further discuss the significance of this comparison between the relative rates at which James and Dewey mention others in (Klein Forthcoming).

How the Study Was Performed

On a MacBook pro (running Big Sur), I downloaded the Stanford Named Entity Recognizer version 4.2.0 (package date November 17, 2020) to my local machine. I opened the NER using an included GUI (I used the file ner-gui.command), and then loaded the 3-class classifier (from the file english.all.3class.distsim.crf.ser.gz).

For each book, I copied the full text from the appropriate text file directly into the GUI window, and then ran the NER. The results appeared in a terminal window. I copied the results from the terminal into an Excel spreadsheet.[1] I then used Excel's text-to-columns wizard, using the ":"

[1] A useful tutorial for running this kind of software is here: https://workbook.craftingdigitalhistory.ca/supporting%20materials/ner/ .

delineator, to break the results into two columns, one with the tag name (either LOCATION, ORGANIZATION, or PERSON), the other with the tagged word or phrase.

I then sorted alphabetically and cleaned the data (see below). Finally, I added columns that trim leading or trailing spaces of each word or phrase tagged as PERSON, then tallied the number of appearances in both the book and in the corpus. The formulae are preserved in the spreadsheets.

The study was later repeated for accuracy with the same version of NER running on Mac's Monterey platform (12.12.1). The results published here are those produced during the second, Monterey-based study.

How the Data Were Cleaned

The data were lightly but not systematically cleaned using the following procedure.

I sorted the results alphabetically, and then manually skimmed through all the results attempting to find and fix any words or phrases that were either erroneously tagged as a PERSON or erroneously not tagged as a PERSON. For example, in Dewey's *Democracy and Education*, the word "Hegel" is erroneously tagged as a LOCATION. I hand-corrected that entry (and numerous similar entries) to the PERSON tag. And in James's *Pragmatism*, the German word "Wie" was erroneously tagged as referring to a PERSON; I removed that tag (James frequently used German and French phrases in his English prose, and NER sometimes mistakenly tags such words). Finally, I deleted all the LOCATION and ORGANIZATION tags, leaving only the corrected PERSON tags.

NER does miss some name references, occasionally. For example, NER identifies 24 instances of the name "Lange" in James's *Principles*, whereas one can use a case-sensitive word search to see that *Principles* in fact contains 28 instances of this name. The results tabulated here only include names positively identified by the NER—I did not supply any missing instances (such as the four overlooked "Lange" instances in the *Principles*). Missed names do not undermine my comparative study on the assumption that names are likely to be undercounted at a similar rate in both the James and Dewey corpora.

Some PERSON tags are connected to fictional or mythological characters, or characters from religious texts such as the bible. While I did not attempt to purge these names when NER tagged them as PERSON, I also did not affirmatively change a tag to PERSON when I could tell that a fictional or mythological character had been erroneously tagged as ORGANIZATION or LOCATION. For example, "Ophelia" was erroneously tagged in James's *Principles of Psychology* with ORGANIZATION, and I did not include that instance in my count of PERSON tags.

I characterize this form of data correction as thorough but not systematic. Between the two corpora there are thousands of lines of results (for example, NER assigns over 2,300 tags to James's *Principles of Psychology* alone, including ORGANIZATION, LOCATION, and PERSON tags). It is impossible *deeply* to review every word without many hours of manual labor or a research team (which I did not employ). The problem is not just the sheer volume of

tags, all of which I read and reviewed; it is that scores of potential errors would require further research to check. For example, in the results for *Democracy and Education* the word "Pestalozzi" was tagged as an ORGANIZATION. A quick web search reveals that this is the name of an 18th century Swiss pedagogue, who I presume is the person to whom Dewey actually referred. That error was corrected, but I could not have systematically caught every such error. There are enough unfamiliar names, organizations, and places mentioned in my two corpora (unfamiliar even to a reader like me who has a solid historical understanding of these authors) that it would take a substantial research team to systematically check every tag.

What is more, even familiar names present challenges. Is "Berkeley" being used to refer to a location or a person? The NER routinely assigned this word the LOCATION tag erroneously. But one must check the context if one wants to ensure accuracy, which I did for instances of "Berkeley" in the *Pragmatism* book (since James had initially presented some of that material in Berkeley, California). I checked the context for many other instances where I thought I could quickly resolve ambiguity, but exhaustiveness is not possible here. These kinds of issues no doubt create some distortion throughout my data in ways my manual skimming of the material did not catch.

Those limitations, however, do not undermine my study's capacity to help us make an informed *comparison* of the frequencies with which persons are mentioned in each corpus, which is my central purpose. Again, we can reasonably assume that the error rate is similar between the two corpora. If that assumption is correct, then taken together, these results should give us a reliable model of relative frequencies of mentions of persons in each corpus. But we should not be as confident that absolute values of person-mentions are precise, for each book or each author.

Finally, I included formulae in my Excel sheets to count instances of each word or phrase tagged with PERSON so we can quickly see how many times the tagged entity appears both in the book in question and in the corpus at large. But I did not attempt to normalize names. For example, the NER tagged "William James" with PERSON in Dewey's *Reconstruction and Philosophy*, but simply tagged "James" with PERSON in Dewey's *How We Think*. Even though these references are presumably to the same person, Excel's countif function tallies them separately. I did not attempt to associate these or any other sets of names together as all referring to one person, since counting of specific names (while interesting) was not my central focus for this study.

The counting functionality suggests further avenues for research, one of which can be illustrated by considering occurrences of the name "Lange." There are no fewer than five different persons with this surname mentioned in the *Principles*: F. A. Lange (PP 40 – 41), Ludwig Lange (PP 99), Nicolai Lange (PP 420 – 421), Karl Lange (PP 751), and Carl Lange (PP 1059 *ff.*). But if one wanted to count only the references to Carl Lange, say, further work would be needed to normalize those name references. For, sometimes James simply refers to "Lange," sometimes to "C. Lange," sometimes to "Herr Lange," and so on, and the NER by itself is not able to group such references in the way a human reader would.

Thus, while we can be more confident about frequencies of references to persons with unique names like "Aristotle," we cannot be as confident about reference rates when it comes to persons with more common names like "Lange," at least not without further normalization work. For a

promising solution to the problem of normalizing author names in the context of a large-scale citation analysis, see (Engelen et al. Forthcoming).

Works Cited

Engelen, Jan, Sander Verhaegh, Loura Collignon, and Gurpreet Pannu. Forthcoming. "A Bibliometric Analysis of the Cognitive Turn in Psychology." *Perspectives on Science*.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*: 363-70.

Klein, Alexander. Forthcoming. "Introduction." In *Oxford Handbook of William James*, edited by Alexander Klein. New York: Oxford University Press.