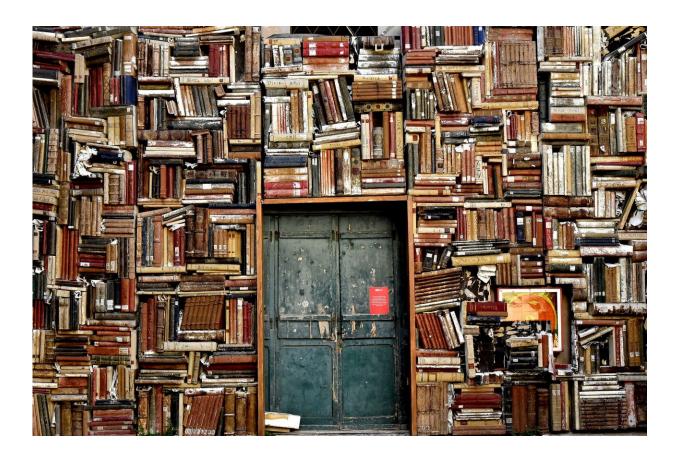# What 26,000 books reveal when it comes to learning language

October 25 2019, by Bert Gambini

Credit: CC0 Public Domain

What can reading 26,000 books tell researchers about how language environment affects language behavior? Brendan T. Johns, an assistant professor of communicative disorders and sciences in the University at

Buffalo's College of Arts and Sciences, has some answers that are helping to inform questions ranging from how we use and process language to better understanding the development of Alzheimer's disease.

But let's be clear: Johns didn't read all of those books. He's an expert in computational cognitive science who has published a computational modeling study that suggests our experience and interaction with specific learning environments, like the characteristics of what we read, leads to differences in language behavior that were once attributed to differences in cognition.

"Previously in linguistics it was assumed a lot of our ability to use language was instinctual and that our environmental experience lacked the depth necessary to fully acquire the necessary skills," says Johns. "The models that we're developing today have us questioning those earlier conclusions. Environment does appear to be shaping behavior."

Johns' findings, with his co-author, Randall K. Jamieson, a professor in the University of Manitoba's Department of Psychology, appear in the journal *Behavior Research Methods*.

Advances in natural language processing and computational resources allow researchers like Johns and Jamieson to examine once intractable questions.

The models, called distributional models, serve as analogies to the human language learning process. The 26,000 books that support the analysis of this research come from 3,000 different authors (about 2,000 from the U.S. and roughly 500 from the U.K.) who used over 1.3 billion total words.

George Bernard Shaw is often credited with saying Britain and America

are two countries separated by a common language. But the languages are not identical, and in order to establish and represent potential cultural differences, the researchers considered where each of the 26,000 books was located in both time (when the author was born) and place (where the book was published).

With that information established, the researchers analyzed data from 10 different studies involving more than 1,000 participants, using multiple psycholinguistic tasks.

"The question this paper tries to answer is, 'If we train a model with similar materials that someone in the U.K. might have read versus what someone in the U.S. might have read, will they become more like these people?'" says Johns. "We found that the environment people are embedded in seems to shape their behavior."

The culture-specific books in this study explain much of the variance in the data, according to Johns.

"It's a huge benefit to have a culture-specific corpus, and an even greater benefit to have a time-specific corpus," says Johns. "The differences we find in language environment and behavior as a function of time and place is what we call the 'selective reading hypothesis.'"

Using these machine-learning approaches demonstrates the richly informative nature of these environments, and Johns has been working toward building machine-learning frameworks to optimize education. This latest paper shows how you can take a person's language behavior and estimate the types of materials they've read.

"We want to take someone's past experience with language and develop a model of what that person knows," says Johns. "That lets us identify which information can maximize that person's learning potential."

But Johns also studies clinical populations, and his work with Alzheimer's patients has him thinking about how to apply his models to potentially help people at risk of developing the disease.

He says some people show slight memory loss without other indications of cognitive decline. These patients with mild cognitive impairment have a 10-15% chance of being diagnosed with Alzheimer's in any given year, compared to 2% of the general population over age 65.

"We're finding that people who go on to develop Alzheimer's across time are showing specific types of language loss and production where they seem to be losing long-distance semantic associations between words, as well as low-frequency words," he says. "Can we develop tasks and stimuli that will allow that group to retain their language ability for longer, or develop a more personalized assessment to understand what type of information they're losing in their cognitive system?

"This research program has the potential to inform these important questions."

Provided by University at Buffalo

provided for information purposes only.