What the Rise of Large Datasets Means for Psycholinguistics

January 20, 2016

The ability to crowdsource data from large groups and the rise of Big Data have helped advance many different areas of psychological research. The field of psycholinguistics — the study of the psychology behind the acquisition, use, production, and comprehension of language — is one of those areas. Such is the importance of Big Data to the field that it was the subject of a special issue, edited by Emmanuel Keuleers (Ghent University, Belgium) and APS Fellow David A. Balota (Washington University in St. Louis, USA) and published in a 2015 issue of *The Quarterly Journal of Experimental Psychology*.

Words are often the main focus of linguistic studies, and variables unique to each word — such as length, pronunciation, frequency, concreteness, and valence — influence how people process and respond to each word. Large datasets that examine these factors allow psychological scientists to understand how each variable affects language processing and recognition, enabling researchers to control for these variables — when not the focus of interest — in their own studies.

As early as the 1940s, researchers created databases chronicling characteristics of words such as their familiarity and vividness. These studies were laboratory-based and were considered large for their time, with several hundred words being examined. As technology has changed to lab-centered, computer-based data collection — and now to Internet-based intake — researchers are more easily able to amass data on tens of thousands of stimuli from thousands of participants; thus, psychological scientists have now been able to create databases examining a wide range of language characteristics such as subjective familiarity ratings, meaningfulness, age of acquisition, valence, arousal, concreteness, and dominance of words. The data from these large-scale and crowdsourcing studies can be used to create norms for words that eventually may be used as stimuli for other researchers.

Such large datasets also can be used to test novel variables, to create predictive computational models of language processing, and to help researchers understand how words gain meaning from the words that surround them and from the context provided by larger chunks of text.

The expansion of these types of datasets brings with it many benefits but also potential methodological concerns. For example, how might one go about replicating a megastudy, and does participant fatigue during data collection reduce the reliability of megastudy findings? Psychological scientists are addressing these questions, with many developing and applying innovative techniques to address the reliability and replicability of these types of studies.

This special issue highlights the utility of large-scale datasets for the field of psycholinguistics and shines a light on researchers who are advancing their field by creating new linguistic databases, utilizing such datasets to better understand the way we process language, and tackling methodological issues that arise with the expansion and application of these techniques.

Reference

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology, 68,* 1457–1468. doi:10.1080/17470218.2015.1051065