The Continued Importance of Theory:

Lessons from Big Data Approaches to Language

Brendan T. Johns
University at Buffalo

Randall K. Jamieson
University of Manitoba

Michael N. Jones
Indiana University, Bloomington

Correspondence
Dr. Brendan Johns
Dept. of Communicative Disorders and Sciences
122 Cary Hall
University at Buffalo
Buffalo, NY, 14214

Email: btjohns@buffalo.edu
Phone: (716) 829-2797
Fax: (716) 829-3979

Big data approaches to cognition have become increasingly popular, coinciding with the continued collection and curation of extremely large collections of human behavior (Jones, 2017). The study of natural language has been particularly impacted by the growth in this area, due to the rise of large and realistic corpora of text and the corresponding development of cognitive models that can learn from this text.

It is now possible to build large-scale computational models of language, train those models with a similar amount of linguistic experience to which an average human may have been exposed to, and determine if the models have extracted knowledge comparable to the average human. We gain insights into learning processes by determining how closely the model's behavior maps onto empirical data collected from human subjects (Landauer & Dumais, 1997). Or, we can use the representations extracted from the models as the basis for modelling other cognitive processes, including lexical organization (Hsiao & Nation, 2018; Jones, Johns, & Recchia, 2012; for a review, see Jones, Dye, & Johns, 2017), episodic memory (Johns, Jones & Mewhort, 2012; Mewhort, Shabahang, & Franklin, 2018), lexical-perceptual integration (Andrews, Vigliocco, & Vinson, 2009; Johns & Jones, 2012; Lazaridou, Marelli, & Baroni, 2017), decision (Bhatia & Stewart, 2018), and sentence processing (Johns & Jones, 2015). Additionally, the development of big data approaches to cognition has led to significant applied solutions in cognitive science, such as determining the changes occurring in lexical semantic memory during aging (Taler, Johns, & Jones, in press) or in the behavior of patients who are developing a memory disorder (Johns et al., 2018). Additionally, there are many applied uses for these models, such as automated essay grading (see Jones & Dye, 2018 for a review). Theoretically, this research area has demonstrated the large and systematic connection between

the natural language environment and human lexical behavior (Landauer & Dumais, 1997; Johns, Jones, & Mewhort, in press; Jones & Mewhort, 2007).

The insights offered by big data approaches to natural language did not emerge in a vacuum. Much of current theory emerging from big data approaches to cognition mimics early work in the cognitive sciences that called for a systematic evaluation of the connection between human behavior and the environments that humans occupy (Estes, 1956, 1975; Simon, 1956, 1969). Specifically, Herbert Simon (1956, 1969) proposed that understanding cognition requires an examination and understanding of the organism, its environment, and the interaction of the two: "the apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves." (Simon, 1969, p. 53) Earlier, William Estes (1955, p. 145) proposed that theories of behavior should shift, "the burden of explanation from hypothesized processes in the organism to statistical properties of environmental events."

As an example of the importance of understanding the external environment in order to understand behavior, Simon (1969) describes an ant walking on a beach. Although the path that the ant takes seems complicated, the complexity in the ant's behavior reflects a series of simple local adjustments to manoeuvre around the obstacles in its way. If one examines the ant's path without considering its environment, one might misattribute the complexity of the path to the ant rather than the environment.

Big data approaches to natural language heed Estes' and Simon's warnings by acknowledging and quantifying the natural language environment that people are embedded in. With the advance of big data, it is no longer necessary to blindly approximate the information structures to which people are exposed: the information can be directly estimated. Indeed, research has shown that the approximations that have been made in the past have not been

accurate, such as in representational assumptions that were made in the cognitive modeling of memory (Johns & Jones, 2010). The availability of large-scale data sources means that making these assumptions is no longer necessary. For example, researchers can estimate the frequency with which different syntactic constructions occur and determine the effect of that exposure on people's behavior (e.g., Reali & Christiansen, 2007). Or, properties of texts can be examined to determine how language changes by the demographic characteristics of authors in order to examine how lexical experience changes lexical behavior (Johns & Jamieson, in press). More generally, advances in big data approaches to cognitive science have allowed for a systematic analysis of the connection between the statistical structure of the environment and human behavior, especially in the study of how people learn, organize, and use natural language (Jones, Willits, & Dennis, 2015; Jones, Dye, & Johns, 2017; Johns, Jones, & Mewhort, in press; Johns, Mewhort, & Jones, in press).

Two of the best examples of how big data is being used in the cognitive sciences are given by the fields of lexical organization and lexical semantics. Lexical organization is the problem of how words are stored and retrieved in the mental lexicon. Early in the investigation of the problem, researchers focused on the impact that environmental variables had on word recognition. The first and still widely used variable is word frequency in written language (e.g., Kucera & Francis, 1967). Word frequency effects are ubiquitous across studies on language processing. For example, high-frequency words are easier to process than low-frequency words (e.g. Broadbent, 1967; Forster & Chambers, 1973; for a recent review, see Brysbaert, Mandera, & Keuleers, 2018). Due to these findings, word frequency has become a central component in models of lexical access (e.g., Goldinger, 1998; Murray & Forster, 2004), and is a standard variable used to control and select stimuli.

For the purposes of this article, the importance of word frequency is how it is calculated. Initially, the values were calculated from the analysis of small corpora, for example, the widely used Kucera and Francis (1967) word counts derived from the Brown corpus of one million words. However, the advent of the internet and powerful computers has allowed better, larger, and more diverse sources of language to be assembled, such as school-age textbooks (Landauer & Dumais, 1997), online encyclopaedias (Shaoul & Westbury, 2010), television and movie subtitles (Brysbaert & New, 2009), social media (Herdağdelen & Marelli, 2017), crowdsourced dictionaries (Johns, 2019), and fiction and non-fiction books (Johns & Jamieson, 2018; Johns, Jones, & Mewhort, in press), among many others. The collection of these different language sources has enabled a closer correspondence between the language that people experience and how language is stored and retrieved. Additionally, it has led to the ability to quantify aspects of our language environment to examine issues such as gender or racial bias (e.g., Caliskan, Bryson, & Narayanan, 2017; Johns & Dye, 2019). As detailed later in this article, the availability of large corpora has also enabled the development on new models of lexical strength (e.g., Adelman, Brown, & Quesada, 2006; Hoffman, Lambon Ralph, & Rogers, 2013; Jones, Johns, & Recchia, 2012), allowing for substantial theoretical growth in this research area.

Related to the rise of new language sources, the field of lexical semantics has also seen rapid change due to both the growing availability of algorithms that can learn from these very large sources of language. Models of this type are referred to as distributional models of semantic memory and propose that word meaning can be inferred from the patterns in which words are used (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov, et al., 2013; Jamieson, Avery, Johns, & Jones, 2018). Although the models differ in significant ways—typically designed to explain particular aspects of linguistic

behavior—they all stem from a similar theoretical basis: simple learning mechanisms applied to language experience establish a sufficient basis for the acquisition of word meanings (Jones, et al., 2015). In short, they all propose an account of language learning in which a word's meaning derives from the company it keeps—consistent with a more general approach dating back to Wittgenstein (1953) who stated (p. 43) that, "The meaning of a word is its use in the language."

More recently, cognitive scientists have leveraged the semantic representations derived from corpus-based models to make sense of behavior as well as linguistic knowledge (e.g. Chubala, Johns, Jamieson, & Mewhort, 2016; Hills, Jones, & Todd, 2012; Hoffman, Lambon Ralph, & Rogers, 2013; Johns & Jones, 2015; Johns, Jones, & Mewhort, 2012, 2019; Taler, Johns, Sheppard, Young, & Jones, 2013). In those accounts, the language representations derived from a distributional model are imported and used in a theory that presents an account of lexical processing, episodic memory, and decision. This exciting program of research aimed at integrating theories of knowledge representation with models of cognitive processing has pushed the borders of psychological theory by allowing for a more complete approach to theory and model development, as it allows for both the process and representational components of a model to be fully specified. Additionally, it simplifies the modeling framework by allowing assumptions about the structure of language in memory to be greatly reduced (Johns & Jones, 2010; Johns, Mewhort, & Jones, 2017). Furthermore, it allows for the simulation of empirical fields that rely upon the semantic content of linguistic stimuli to manipulate behavior, such as proactive interference (Mewhort, Shabahang, & Franklin, 2018) and false memory (Johns, Jones, & Mewhort, 2012, 2014).

Another exciting aspect of corpus-based models is that they enable an examination of model performance at the item level. For example, using a distributional model of semantics, one

could take the similarity between the words *dog* and *wolf*, and determine how closely that similarity maps onto people's similarity judgements for those same words. Ideally, this would be analyzed over hundreds or thousands of word pairs, enabling a systematic analysis of how closely a model's representation of word meaning maps onto human behaviour. This would allow for a determination as to whether the learning mechanism that a model employs successfully captures the required lexical behavior. Fortunately, the field was forward-looking enough and a number of large-scale data collection projects have enabled just this type of analysis, meaning that the impact of big data on the cognitive sciences has both a theoretical and empirical component.

Mega Datasets of Human Behavior

Both lexical organization and lexical semantics have been targets for large-scale data collection, an exercise that has had a major impact on both fields. In the field of lexical organization, the English lexicon project (ELP; Balota, et al., 2007) has allowed the field to account for lexical behaviors at an item level. The English lexicon project collected lexical decision and naming reaction time performance for over 40,000 words from over 800 subjects at sites around the United States. The major impact of this work, apart from providing empiricists with a more refined ability to control stimuli, is that it has allowed a number of new models on lexical organization to be tested at the item-level (e.g. Adelman, Brown, & Quesada, 2006; Jones, Johns, & Recchia, 2012). Indeed, the ELP proved so successful that parallel projects have been conducted for a number of different languages and dialects, including British English (Keuleers, Lacey, Rastle, & Brysbaert, 2012), French (Ferrand et al., 2010), Dutch (Brysbaert, Stevens, Mandera, & Keuleers, 2016), Chinese (Tse et al., 2017), and Malay (Yap, Liow, Jalil, & Faizil, 2010). The availability of lexical information across so many different languages is an

underutilized resource, but there are some ways these datasets have been exploited, such as demonstrating that lexical organization models generalize across different languages (Jones, Dye, & Johns, 2017). Other studies have expanded on the ELP by publishing related but different measures of lexical processing (e.g., the semantic decision project; Pexman, Heard, Lloyd, & Yap, 2017). Overall, the existence of so many different lexicon projects demonstrates the use that researchers have found from the original ELP, and the promise that many researchers recognize in conducting a large-scale item-level analyses of lexical behavior.

Lexical semantics has an even older and arguably richer history of using mega datasets to examine item-level properties of word meaning. Nelson, McEvoy, and Schreiber (2004) published the University of South Florida free association norms, which contains association data for 72,000 word pairs and has been used to both examine the nature of free association (Nelson, McEvoy, & Dennis, 2000) and develop new models of lexical semantics (e.g., Griffiths et al., 2007).

Similarly, McRae, Cree, Seidenberg, and McNorgran (2005) published a large set of feature production norms, where subjects were asked to generate as many defining features about a noun as they could. This method is unique in defining semantic representation because it collects in formation on people's mental representations that might not be encoded directly in text (Cree & McRae, 2003) and, thereby, has served an important role for the development of models that integrate lexical and perceptually grounded representations (e.g. Andrews, Vigliocco, & Vinson, 2009; Johns & Jones, 2012; Riordan & Jones, 2011). Vinson and Vigliocco (2008) published a similar set of norms, but with events included along with objects.

Taking a similar approach to the ELP, the Semantic Priming Project (Hutchison, et al., 2013) collected semantic priming data for both lexical decision and naming time for thousands of

words across hundreds of subjects. Although the semantic priming project has not been the target for as much theoretical work as the ELP, it has provided a good deal of information about the nature of semantic priming, such as the impressive range of individual differences in semantic associations (Yap, Hutchison, & Tan, 2017).

This quick summation of available large-scale datasets demonstrates the importance that researchers are placing on item-level analyses of human behavior. Indeed, the item-level data available to researchers spans many areas across the study of language and cognition, including such diverse data types as idiomatic processing (Bulkes & Tanner, 2017), word associations (De Deyne, Navarro, Perfors, Brysbaert, & Storms, in press), taboo words (Roest, Visser, & Zeelenberg, 2018), modality norms (Lynott & Connell, 2009, 2013), humor (Engelthaler & Hills, 2018), and body-object interaction ratings (Bennett, Burnett, Siakaluk, & Pexman, 2011; Tillotson, Siakaluk, & Pexman, 2008; Pexman, Muraki, Sidhu, Siakaluk, & Yap, 2019). A particularly noteworthy project is the Canadian Longitudinal Study of Aging (CLSA; Raina, et al., 2009; Tuokko, et al., 2019) which is collecting both demographic and behavioral data on a cohort of 50,000 participants across the aging spectrum on a yearly basis, and has become a target for distributional models of semantics (Taler, et al., in press).

The multi-lab effort to collect large sources of data across different tasks signals the growing belief in the power that item-level analysis holds for developing and evaluating new theories of cognition, and especially the cognitive processes underlying language and memory. Critically, it not only offers new resources with which to control experiments, but also signals new pathways to theory development. Particularly, large-scale data collection is complementary to the rise of corpus-based modeling, as the large scale analysis of the language environment allows for the training of cognitive models at a scale of language experience that is comparable

to what a person might encounter. Secondly, empirical big data allows for hundreds or thousands of word-specific data points to be used in model tests. However, this form of theory development offers a significant divergence from the traditional hypothetico-deductive method that is standard practice in the psychological sciences.

Classic and Big Data Approaches to Theory Development

In the field of experimental psychology, and particularly in the cognitive sciences, the hypothetico-deductive method has served as the dominant paradigm for theory development (Laudan, 1981; Hayes, 2000). In practice, the method involves using a theory to produce a priori predictions about the influence of a particular experimental manipulation on human behavior. These predictions are then tested in controlled target experiments designed to test the prediction by a criterion of falsification. To the extent that the data match the a priori theoretical predictions, the theory remains tenable. When data contradict the theory, the data prevail and the theory is ruled-out or, more often, is modified to accommodate the discrepancy. This leads to an iterative process in which a theory is proposed, predictions are generated, data are collected, and the theory is evaluated. The hope is that across experiments more powerful and robust theories of human cognition are developed. Clearly, this is an oversimplification of the complexities associated with modern scientific work, but experimental psychology does hold to the foundations of this approach (cf. Roozeboom, 1999).

In contrast to the hypothetico-deductive approach emphasized in experimental psychology, theory development in big data relies more on abduction, a form of reasoning not without its champions in the methodology literature (see Haig, 2005, in press; Roozeboom, 1999; this approach is also popular in grounded approaches to psychological theory, see Rennie, Phillips, & Quartaro, 1988). As Haig (2005, pp. 372-373) states, abductive reasoning

"…involves reasoning from phenomena, understood as presumed effects, to their theoretical explanation in terms of underlying causal mechanisms." That is, the data are first collected, and then theories of that data are formed from knowledge in the domain. These theories can then be contrasted with other theories in terms of their goodness of fit to relevant data, and can be subject to the same empirical testing that traditional approaches employ. Theory development thus emerges from data that are already collected, not from necessity of testing a hypothesis. Much of the large-scale data collection projects outlined above were collected with this strategy in mind – once the data is available to researchers, new theories that can explain the variance in the data are constructed. That is, serious resources are being devoted not to test a given theory, but with the hope that new theories can emerge from collected data.

As an illustration of the success of the abductive approach to cognition, consider the development of contextual diversity and semantic diversity models of lexical organization (for a recent review of this area, see Jones, Dye, & Johns, 2017). The first work in this area was an article put forth by Adelman, Brown, and Quesada (2006) who demonstrated that a contextual diversity (CD) count provided a better fit to lexical decision and naming time (taken mainly from the ELP) data than a word frequency (WF) count. A contextual diversity count was operationalized as the number of different documents that a word occurs in across a corpus, not just the number of times that a word occurs in the corpus. The advantage of a context count has been replicated across a number of corpora and datasets (see Adelman & Brown, 2008; Brysbaert & New, 2009).

Jones, Johns, & Recchia (2012) agreed with many of the proposals of Adelman, et al. (2006), but hypothesized that the semantic construction of a context should also be important in building a word's strength in memory (see Hoffman, Lambon Ralph, & Rogers, 2013, for a

similar proposal). Specifically, they hypothesized that words that occur in many different contexts (e.g., *occasion*) should be stored in memory more strongly than words that appear in mainly redundant contexts (e.g., *molecule*). To test this hypothesis, Jones, et al. developed a distributional model that learns from the semantic context that a word occurs in, entitled the Semantic Distinctiveness Model (SDM). Crucially, the SDM employs an expectancy-congruency mechanism such that the strength that a given context is stored in memory for a word is based on how unique that context is compared to the past contexts that the word occurred in, with more unique contexts being given a stronger representation in memory, relative to words that occur in more redundant contexts. Using this learning mechanism, the SDM provided better fits than CD and WF measures to lexical decision and naming time data from the ELP. The analysis demonstrates the importance of acknowledging and considering the semantic rather than text-derived definition of a context when explaining patterns of lexical organization.

Furthermore, it has been demonstrated that the SDM provided an advantage over WF and CD in spoken word recognition (Johns, Gruenenfelder, Pisoni, & Jones, 2012), could account for results in a controlled natural language learning (Johns, Dye, & Jones, 2016), and that the model generalized to alternative populations, such as bilinguals and older adults (Johns, Sheppard, Jones, & Taler, 2016). That is, several models (Adelman, et al., 2006; Jones, et al., 2012) were derived to explain patterns in the ELP mega-data set, which were then contrasted and validated using a mix of targeted experimentation and abductive fits. The result of this research program was a new and powerful model of lexical organization, the development of which would have been impossible without access to large-scale sets of data prior to model development. The exercise shows how abductive and deductive reasoning can be combined productively and carefully: abductive reasoning is used to develop insightful new models that are, then, tested in

targeted experiments designed to evaluate and refine those models, through the use of abductive reasoning. The ultimate goal of abduction in the scientific reasoning process is inference to the best explanation, that is, the model that can explain the relevant data better than any known alternative.

The mechanisms of the SDM are rooted in the structure of the natural language environment. That is, the performance of the model is dependent on information gleaned from very large text databases. This means that the explanatory power of the model comes from the systematic connection between the structure of the language environment and the patterns contained in large sets of behavioral data. As stated previously, corpus-based models are uniquely suitable to the abductive approach to model development, since the outputs of this models can take place at the group (Johns, et al., 2016), individual (Johns & Jamieson, 2018), or item-level (Jones, et al., 2012), meaning that the plausibility and power of these theories can be tested and validated with both abductive and deductive empirical approaches.

However, corpus-based models are not the only type of theories built by abductive reasoning. Other researchers have been using large-scale collections of data to explain variance in other large-scale sets of data. We believe this approach represents an error in reasoning, with Jones, Hills, and Todd (2015) having labeled these approaches "Turk Problems."

Turk Problems

The Turk was a chess-playing machine developed in 1770 by Wolfgang von Kempelen to impress the Empress of Austria. It was purported to be an automaton that could play a realistic game of chess against a human opponent. However, in reality, the machine was an illusion – there was a human chess master housed inside the machine that controlled the operation of the Turk.

Given this illusion, it is clear that in order to understand how the Turk works it would be first necessary to understand how a human plays chess. Jones et al. (2015) used this analogy to describe models that utilize human behavioral data (e.g., free association values) as an underlying representation. In cognitive modeling, a Turk problem arises when a model's representation is derived from human behavioral data, hiding the complexity of the model within the model's representation.

To understand the behavior of a model that uses a representation derived from human behavior, it is necessary to understand the data that the model is using as its representation. Model behavior is not independent of its choice in representation, as Hummel and Holyoak (2003; p. 247) note, "All models are sensitive to their representations, so the choice of representation is among the most powerful wild cards at the modeller's disposal." In some cases, the underlying data used to form a representation can be more complicated than the data that is being modeled. This means that the complexity of a model is significantly undervalued, and that the understanding derived from a model is muddled at best.

We believe that Turk problems represent an error in abductive reasoning. The goal of abductive reasoning in this context is to find patterns in data that point to better theories, such as Adelman et al. (2006) did to determine that contextual diversity provides a better explanation of lexical behaviour than word frequency. There is knowledge gained from such an analysis about the nature of the cognition under question, and the results spurned new ideas for additional empirical and computational research (e.g., Jones et al., 2012). However, demonstrating that one kind of data provides a good fit to another kind of data offers little theoretical progress, unless one kind of data is already completely understood (which is rarely the case in the cognitive sciences). As an example, consider free association norms.

Free association norms. Semantic verbal fluency is a common task used in both clinical and theoretical settings to explore memory search and semantic memory performance (see Taler & Phillips, 2008, for a review). The most common version of this task involves naming as many items from a category as possible (e.g., animals) within a set time limit (typically one minute). Traditionally, behavior in this task is assessed by a count of how many category items a subject produced. However, Hills, Jones, & Todd (2012) used a memory search model over a representation of word knowledge from a distributional model of semantics (i.e., BEAGLE; Jones & Mewhort, 2007) to confirm that critical features in peoples' pattern of recall was consistent with critical features in animals' food foraging behaviour. This work shed light on the generality of cognitive search mechanisms across species (Hills, Todd, & Jones, 2015), and has been adapted to explore memory search in bilinguals (Taler, Johns, Young, Sheppard, & Jones, 2013) and in patients who are developing cognitive impairment (Johns et al., 2018).

Abbott, Austerweil, and Griffiths (2015) questioned this work and proposed that a random walk model (a model type that has had success in the past; e.g., Griffiths, Steyvers, Firl, 2007) could provide an equivalent explanation of verbal fluency behavior. However, their model did not capture the appropriate patterns when using the BEAGLE representation, but could when it used a semantic network derived from the free association norms of Nelson et al. (2004). Free association norms had been used previously to drive models of semantics (e.g., Steyvers, Shiffrin, & Nelson, 2004).

As Jones et al. (2015) point out, the use of free association norms subverts traditional accounting of model complexity. The representations derived from BEAGLE come from an articulate and well understood computational model applied to stable statistical properties of the natural language environment. In contrast, free association data represents patterns of human

behavior from a cognitive process that is not well understood. However, by a simple accounting of the parameter space of the two models, the approach of Abbot, et al. (2015) could be seen as equal to or better than the approach of Hills, et al. (2012), when representational complexity is not considered.

If we accept Abbott et al.'s (2015) account, there is an additional question of what has actually been understood about verbal fluency. If verbal fluency performance is no more than a random walk over free association strength, then it is necessary to first understand the cognitive underpinnings of free association. Given that computational models of semantics give a poor accounting to item-level effects in free association (see Maki, McKinley, & Thompson, 2004), it is difficult to accept that this is a well understood data type. Thus, the model has simply shifted the theoretical goals from understanding verbal fluency to understanding free association. However, verbal fluency performance is quite well captured by searching mechanisms over representations derived from distributional models (e.g. Hills, et al., 2012; Johns, et al., 2018; Taler, et al., 2013), suggesting that verbal fluency performance can be accounted for without a need to first understand the cognitive processes that underpin free association.

Internal versus external theories of language. Like Abbot, et al. (2015), De Deyne, Perfors, and Navarro (2016) have proposed that a model based on word association data provides a better fit than distribution models to peoples' lexical behaviour – similar to the notion of using free association data to derive semantic representations (Steyvers, et al., 2004). Specifically, they propose that internal models of language provide a better account than external models of language. De Deyne, et al. (2016) define an external model of language as being one that learns from the natural language environment, such as distributional models. In contrast, internal models of language are models that learns language from the knowledge of speakers of that

language. Their internal language models were derived from a very large set of word association values, where over 80,000 subjects were given a word association task (De Deyne, et al., 2016; De Deyne, Navarro, & Storms, 2013; De Deyne, Navarro, Perfors, & Storms, 2016). The specific word association task used in these studies asked subjects to produce three associates to a given cue word.

To compare the performance of external and internal language models, De Deyne, et al. (2016) contrasted the performance of these model types on word similarity and relatedness tests, where subjects were asked to rate the similarity (or relatedness) of a word pair, a standard data type within computational linguistics (see Finkelstein, et al., 2001). De Deyne et al. show that similarity metrics derived from their word association data outperform a variety of different distributional model types.

The conceptualization of internal versus external models of language effectively characterizes a Turk problem. A word similarity task almost certainly relies on similar cognitive mechanisms as a word association task. Just as explaining lexical decision response times with naming response time would provide very little theoretical insight into the nature of lexical retrieval, the finding that word association data provides a good fit to word similarity data also provides very little theoretical insight into lexical semantics, other than that the two tasks bear some relation. To understand how humans judge the similarity of words in this approach requires an understanding of how humans perform word associations, similar to the problems seen with Abbott et al. (2015).

The goal of distributional models is to explain how people acquire the knowledge that they have (Landauer & Dumais, 1997). From this perspective, an internal model of language is not a competing model to distributional models (i.e., an external model), but instead it is what

this model class was designed to explain (i.e., the knowledge that people have acquired). Thus, it is inappropriate to compare the performance of these model types against one another. Instead, a more productive route would be to determine how well the knowledge gained by external models of language map onto the patterns of knowledge seen in internal models, the original goal of Landauer and Dumais (1997).

Although the distributional models did not perform as well on tests of word similarity and relatedness, they do still provide theoretical insight; by demonstrating that people's semantic similarity judgements are systematically related to the statistical patterns of word occurrence in the natural language environment. However, as De Deyne et al. (2016) show, these models are far from perfect and like all theories need to improved or modified to better fit human lexical behavior. The data collected and used by De Deyne et al. provide a promising pathway to examining the failures of distributional models by determining where the models diverge from word association data on a large scale. However, word association data are still data that need to be understood and explained with theoretical accounts of language processing, not used as a replacement for a theory.

Summary and Conclusions

Big data approaches to cognition have been transformative, as they allow for an examination of human behavior at a level of precision and scale that was not previously possible. It is now possible to propose a cognitive model, train that model with a similar history of language experience that an adult human might have, and test how well that model's behavior maps onto human behavior at the item level. Big data approaches to natural language have been particularly valuable. There have been two main developments that have proved influential on the understanding of language. The first is the collection of large textbases of natural language

and models that can exploit them, such as Latent Semantic Analysis (Landauer & Dumais, 1997). The second is the collection of mega datasets of human behavior, such as the English lexicon project (Balota, et al., 2007). These two developments are interdependent, as large-scale cognitive models must be evaluated at the item-level and those are the precision of data that the mega collections of human behavior provide.

However, theoretical development using big data does not always follow traditional methodologies. Specifically, much of psychological science has used the hypothetico-deductive method of theory development, where theories are used to generate hypotheses about human behavior that are tested in target experiments, and evaluated against the match between prediction and observation. In contrast, much of theory development using big data approaches has proceeded by abduction: data is collected and theories are developed to capture variance in the data. Although abduction is an important part of the scientific process, insights reached by abduction are not equivalent to conclusions reached by deductive experimental verification. For the big data approach to science to succeed, the field needs to translate curious abductive insights into clear deductive conclusions (although others have different opinions about the outcome of abductive reasoning; see Haig, in press).

There are dangers to relying only upon abduction in theory development, however. One of these dangers is the Turk problem (Jones, et al., 2015), where behavioral data is used to explain other behavioral data (e.g., using free association norms to explain verbal fluency). Showing that one type of data predicts another offers no clarity towards theoretical progress. Data are not theory (see Haig, in press, for a different perspective on this issue).

Although this article has mainly focused on two particular problems of cognitive psychology, lexical organization and lexical semantics, the developments in big data has an

increasingly wide-ranging impact in the cognitive sciences. The biggest advantage that corpus-based approaches to cognition offer is that they allow for content to be placed in memorial representations. For example, the classic approach in the computational cognitive modeling of language is to use randomly generated representations of words (Johns & Jones, 2010). This is no longer necessary (see Johns, Mewhort, & Jones, 2017, for a longer discussion of this issue). It is now possible to use a model of distributional semantics (e.g., LSA, BEAGLE, Topics, or a related model) as the representation that can be fed into a process model, allowing for an integration of the knowledge and process components of cognition. Multiple models of cognition have been developed using this integrative approach, such is in episodic memory (Johns, Jones, & Mewhort, 2012; Mewhort, Shabahang, & Franklin, 2018), implicit learning (Chubala, et al., 2016), and decision making (Bhatia & Stewart, 2018). Additionally, the models can be used to determine the underlying cognitive differences in clinical populations, for example in patients with memory disorders (Johns, et al., 2018) or patients with schizophrenia (Minor, Willits, Marggraf, Jones, & Lysaker, in press; Willits, Rubin, Jones, Minor, & Lysaker, in press).

There are a number of challenges facing theoretical development in big data approaches to cognition. One area relevant to the discussion of Turk problems is model complexity. Traditional methods of model comparison use an accounting of a model's parameter space to account for differences in model complexity (e.g., Akaike's Information Criterion; Akaike, 1974). Jones et al. (2015) point out that models that use representations derived from human behavior hide complexity due to the cognitive processes of the subjects used to collect the data inside the representation. There are also complexity issues with distributional models. For example, Recchia and Jones (2011) demonstrate that a simple model of distributional semantics (pointwise mutual information; Bullinaria, & Levy, 2007) can exceed the performance of a more

computational complex model (LSA; Landauer & Dumais, 1997) when the simpler model is given more training materials. However, there are two sources of complexity in this analysis: the computational complexity of the cognitive model, as well as corresponding issues of cognitive plausibility, and the amount of training materials that a model requires to learn from. To continue developing better, more cognitively plausible large-scale models of cognition it will be necessary to develop formal frameworks that can accommodate these different sources of complexity, in order to enable better model comparison techniques. New empirical work, such as the results of Brysbaert, Stevens, Mandera, and Keuleers (2016), which measured the amount of linguistic experience young adults likely have had, provides useful guidelines to perform more cognitively plausible model training.

Big data is a young but maturing field in the cognitive and psychological sciences. Just like any developing field, there are methodological and theoretical challenges to be faced. However, as this chapter outlines, the continued construction of large-scale models of cognition, together with the collection of mega datasets of human behavior, provide a promising foundation for the development of increasingly powerful theories of human behavior.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, 122, 558-569.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision time. *Psychological Science*, 17, 814-823.

Adelman, J. S., & Brown, G. D. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, 115, 214.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116, 463.

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, 179, 71-88.

Balota, D. A., Cortese, M. J., Hutchinson, K. A., Neely, J. H., Nelson, D., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 339, 445-459.

Bennett, S. D., Burnett, A. N., Siakaluk, P. D., & Pexman, P. M. (2011). Imageability and body–object interaction ratings for 599 multisyllabic nouns. *Behavior Research Methods,* 43, 1100-1109.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review, 74*, 1-15.

Brysbaert, M., & New, B. (2009). Moving beyond Kucèra and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977-990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 441.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7, 1116.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27, 45-50.

Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 49, 772-783.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39, 510-526.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356, 183–186.

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to the implicit rule-learning task: Implicit learning of semantic structure. *Quarterly Journal of Experimental Psychology*, *69*, 1049-1055.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132, 163.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45, 480-498.

De Deyne, S., Navarro, D.J., Perfors, A., Brysbaert, M., & Storms, G. (in press). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1861-1870).

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, 145, 1228.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, 50, 1116-1124.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, 62, 369.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, *12*, 263-282.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42, 488-496.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web* (pp. 406-414). ACM.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627-635.

Goldinger, S. D. (1998). Echoes of echoes? An episodic trace theory of lexical access. *Psychological Review, 105*, 251-279.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, 18, 1069-1076.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological methods*, 10, 371.

Haig, B. D. (in press). Big data science: A philosophy of science perspective. In Woo, S. E., Proctor, R., and Tay, L. (Eds.) *Big Data Methods for Psychological Research*: New horizons and Challenges, APA Books.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, 41, 976-995.

Hayes, N. (2000). *Doing psychological research*. Abingdon: Taylor & Francis Group.

Hills, T. T., Jones, M. & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*, 431-440.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, 7, 513-534.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45, 718-730.

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, 103, 114-126.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, 45, 1099-1114.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, 1, 119-136.

Johns, B. T. (2019). Mining a crowdsourced dictionary to understand consistency and preference in word meanings. *Frontiers in Psychology*, 10, 268.

Johns, B. T. & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin and Review, 17*, 662-672.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*, 486-518.

Johns, B. T., & Jones, M. N. (2012). Perceptual Inference through global lexical similarity. *Topics in Cognitive Science*, 4, 103-120.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *Journal of the Acoustical Society of America*, 132, EL74-EL80.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, *69*, 233-251.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, 26, 103-126.

Johns, B. T., Dye, M. W., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, 4, 1214-1220.

Johns, B. T., Sheppard, C., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in lexical organization across aging and bilingualism. *Frontiers in Psychology*, 7, 703-714.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2017). Small worlds and big data: Examining the simplification assumption in cognitive modeling (pp. 227-245). In Jones, M. N. (Ed.), *Big data in cognitive science: From methods to insights*, Taylor & Francis.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (in press). The role of negative information in distributional semantic learning. *Cognitive Science*.

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F. W., & Jones, M. N. (2018). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology*, 72, 117-126.

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance of written language. *Cognitive Science*, 42, 1360-1374.

Johns, B. T. & Dye, M. (2019). Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods*, 51, 1601-1618.

Johns, B. T. & Jamieson, R. K. (in press). The influence of time and place on lexical behavior: A distributional analysis. *Behavior Research Methods*.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology,* 66, 115-124.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, & Griffiths. *Psychological Review*, 122, 570-574.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory (pp. 232-254). In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology*. New York: OUP.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizational principle of the lexicon. In B. Ross (Ed.), *The Psychology of Learning and Motivation*, 67:43.

Jones, M. N. (2017). Developing cognitive theory by mining large-scale naturalistic data. In M. N. Jones (Ed.), *Big Data in Cognitive Science*. New York: Taylor & Francis.

Jones, M. N., & Dye, M. W. (2018). Big data methods for discourse analysis. In Schober, M. F., Rapp, D. N., & Britt, M. A. (Eds.) *Handbook of discourse processes*, 2nd Edition.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44, 287-304.

Kučera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Dartmouth Publishing Group.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Laudan, L. (1981). *Science and hypothesis*. Dordrecht, the Netherlands: Reidel.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41, 677-705.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41, 558-564.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, 45, 516-526.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, 36, 421-431.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37, 547-559.

Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review*, 25, 932-950.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality (pp. 3111-3119). In C. J. C. Burges, L, Bottou, M., Welling, Z., Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26*.

Minor, K. S., Willits, J. A., Marggraf, M. P., Jones, M. N., & Lysaker, P. H. (in press). Measuring disorganized speech in schizophrenia: automated analysis explains variance in cognitive deficits beyond clinician-rated scales. *Psychological Medicine*.

Murray, W. S., & Forster, K. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*, 721–756.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, 28, 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36, 402-407.

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, 49, 407-417.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. Behavior Research Methods, 51, 453-466.

Raina, P. S., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., ... & Wister, A. (2009). The Canadian longitudinal study on aging (CLSA). *Canadian Journal on Aging*, 28, 221-229.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57, 1-23.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41, 647-656.

Rennie, D. L., Phillips, J. R., & Quartaro, G. K. (1988). Grounded theory: A promising approach to conceptualization in psychology? *Canadian Psychology*, 29, 139.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3, 303-345.

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2017). Dutch taboo norms. *Behavior Research Methods*, 50, 630-641.

Rozeboom, W. W. (1999). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Hillsdale, NJ: Erlbaum.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, 42, 393-413.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129-138.

Simon, H. A. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237-249.

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30, 501-556.

Taler, V., Johns, B. T., Young, K., Sheppard, C., & Jones, M. N. (2013). A computational analysis of semantic structure in bilingual verbal fluency performance. *Journal of Memory and Language*, 69, 607-618.

Taler, V., Johns, B. T., & Jones, M. N. (in press). A large scale semantic analysis of verbal fluency across the aging spectrum: Data from the Canadian longitudinal study on aging. *Journal of Gerontology: Psychological Sciences*.

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, 40, 1075-1078.

Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, 49, 1503-1519.

Tuokko, H., Griffith, L. E., Simard, M., Taler, V., O'Connell, M. E., Voll, S., ... & Raina, P. (2019). The Canadian longitudinal study on aging as a platform for exploring cognition in an aging population. *The Clinical Neuropsychologist*, 30, 1385-4046.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40, 183-190.

Willits, J. A., Rubin, T., Jones, M. N., Minor, K. S., & Lysaker, P. H. (in press). Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia. *Schizophrenia Research*.

Wittgenstein, L. (1953). *Philosophical investigations*. John Wiley & Sons.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42, 992-1003.

Yap, M. J., Hutchison, K. A., & Tan, L. C. (2016). Individual differences in semantic priming performance: insights from the Semantic Priming Project. In M. N. Jones (Ed.), *Big Data in Cognitive Science*. New York: Taylor & Francis.