Volume 68 Number 8 August 2015

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY

Published for the Experimental Psychology Society

Editor Marc Brysbaert

SPECIAL ISSUE

MEGASTUDIES, CROWDSOURCING, AND LARGE DATASETS IN PSYCHOLINGUISTICS

Guest Editors
Emmanuel Keuleers and David A. Balota

Routledge
Taylor & Francis Group

The Quarterly Journal of Experimental Psychology

Routledge
Taylor & Francis Group

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: http://www.tandfonline.com/loi/pqje20

Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments

Emmanuel Keuleers & David A. Balota

To cite this article: Emmanuel Keuleers & David A. Balota (2015) Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments, The Quarterly Journal of Experimental Psychology, 68:8, 1457-1468, DOI: 10.1080/17470218.2015.1051065

To link to this article: http://dx.doi.org/10.1080/17470218.2015.1051065

Accepted online: 14 May 2015.

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
Vol. 68, No. 8, 1457–1468, http://dx.doi.org/10.1080/17470218.2015.1051065

Introduction

Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments

Emmanuel Keuleers[1] and David A. Balota[2]

[1]Department of Experimental Psychology, Ghent University, Gent, Belgium
[2]Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

(Received 7 May 2014; accepted 8 May 2015)

This paper introduces and summarizes the special issue on megastudies, crowdsourcing, and large datasets in psycholinguistics. We provide a brief historical overview and show how the papers in this issue have extended the field by compiling new databases and making important theoretical contributions. In addition, we discuss several studies that use text corpora to build distributional semantic models to tackle various interesting problems in psycholinguistics. Finally, as is the case across the papers, we highlight some methodological issues that are brought forth via the analyses of such datasets.

Keywords: Megastudies; Crowdsourcing; Language processing; Corpora; Distributional semantics.

This special issue on *megastudies, crowdsourcing, and large datasets in psycholinguistics* collects the most recent research on a number of interrelated developments: *Megastudies* involve the collection of behavioural data on a large number of linguistic stimuli—now typically in the order of tens of thousands; *crowdsourcing* refers to studies in which data are collected outside of the traditional controlled laboratory settings (e.g., MTurk and/or Smart Devices). Of course, both approaches lead to *large datasets*, a term with which we refer more generally to any large collection of linguistic data produced by humans, such as text corpora, that are used in psycholinguistic research.

The evolution of experimental psycholinguistics is intricately linked to the availability of data on the properties of linguistic stimuli. When these stimuli are words—the linguistic unit targeted in most of the articles in this issue– their length, spelling, syntactic class, and pronunciation are the primary properties that are knowable without access to additional data. Text corpora allow us to count how often and where words occur and to explain and control for how their frequency of occurrence affects word recognition. More importantly, research has shown that word processing is affected by a multitude of variables such as age of acquisition, imageability, concreteness, valence, dominance, arousal, and so on, and it is therefore important not only to have norms available for these measures, but to understand how these variables contribute to processing words. Finally, words do not exist in isolation: Orthographic and phonological neighbourhood measures, word association data, and semantic vector spaces built from text corpora tell us how words relate to each other within and across languages.

Correspondence should be addressed to Emmanuel Keuleers, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, Gent 9000, Belgium. E-mail: emmanuel.keuleers@ugent.be

Testing any nontrivial hypothesis about language processing experimentally depends on these data, and so does developing and testing computational models of language processing or production. Without this information, psycholinguistics, and for that matter any area that uses verbal stimuli (e.g., memory and attention studies), would simply be impossible.

The realization that progress in psycholinguistics depends on access to such data is reflected in the influence of the projects in which these data are made easily accessible. The MRC psycholinguistic database (Coltheart, 1981) was made available in the early 1980s and has subsequently been used by several generations of researchers. Since the early 1990s, CELEX (Baayen, Piepenbrock, & Gulikers, 1995) has provided a wealth of lexical information for Dutch, English, and German, and it can be argued that the relatively large amount of research on Dutch in psycholinguistics was for a large part driven by this development. With the English Lexicon project (ELP; Balota et al., 2007) a radical innovation was introduced. Not only did the project provide information on the lexical characteristics for the more than 40,000 English word types from the Brown Corpus, but it also provided chronometric measurements for those items in lexical decision and naming. This has opened new ways of *doing* psycholinguistics. The wealth of chronometric data available in the ELP and similar databases means that many research questions can now be answered by statistical analysis of already available data. This approach has been extended more recently with the advent of crowdsourcing methods (Dufau et al., 2011; Mason & Suri, 2011; Munro et al., 2010; Schnoebelen & Kuperman, 2010) in which data can be collected for a very larger number of items in a fast and reliable manner.

What all papers in this special issue have in common is that they show how new insights can be gained from the abundance of data that psycholinguists now have access to. Many of the papers introduce new data and methods and show how these data address important theoretical and methodological issues. In discussing these contributions, we focus on the following issues. First, we provide a brief historical overview and show how the papers in this issue have extended the field by compiling new databases and making important theoretical contributions. Second, we discuss several studies that use corpora to build distributional semantic models to tackle several interesting problems in psycholinguistics. Finally, as is the case across the papers, we highlight some methodological issues that are brought forth via the analyses of such datasets.

MEGASTUDY DATABASES: A BRIEF OVERVIEW AND NEW INSIGHTS

Databases with subjective ratings, or norms, of lexical characteristics have long been an indispensable part of psycholinguistic investigation. One of the first large lists of norms was collected by Haagen (1949). Questioning the ad hoc judgements of material by experimenters, Haagen published ratings of familiarity and vividness of imagery for more than 400 adjectives, in addition to ratings for synonymity and association of 440 adjective pairs. Twenty years later, a very influential set of norms was published by Paivio, Yuille, and Madigan (1968), who collected ratings for concreteness, imageability (or imagery), and meaningfulness for 925 English nouns. The central insight of these researchers was that there was no reason for words to be rated for every single experiment. By putting some effort in collecting them for a larger number of items, the data would become reuseable, freeing up valuable research time. Quick in abandoning the paper-and-pencil methods, Nusbaum, Pisoni, and Davis (1984) used computer presentation and button box responses to collect subjective familiarity ratings for 20,000 words in the Merriam-Webster pocket dictionary. Balota, Pilotti, and Cortese (2001) reported one of the first web-based subjective familiarity rating studies, in which they had 1590 individuals rate 2938 monosyllabic words. They also found that the familiarity ratings from the web-based participants were strongly correlated with ratings taken from more standard laboratory tasks

($r = .95$ for young adults and $r = .92$ for old adults). Recently, collecting ratings has become the domain of crowdsourcing. Because online platforms can now be used to collect behavioural data, the pool of potential participants is much larger than that for laboratory experiments. Hence, the amount of data that is usually collected is also large. There are now ratings for age of acquistion (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), valence, arousal, and dominance (Warriner, Kuperman, & Brysbaert, 2013) and concreteness (Brysbaert, Warriner, & Kuperman, 2013) for tens of thousands of words.

In hindsight, one can ask why the psychologists who understood the benefit of collecting elicited ratings for a large number of words did not gather chronometric measures for recognition or classification of those words. One possibility is that the reuse of independent variables was considered safe but that recycling a dependent variable did not conform to the idea that formulating a hypothesis must always precede the collection of data in the cycle of scientific investigation. The fundamental idea behind that principle, however, is to prevent a hypothesis being generated *based on* data that are already collected. It is clear to see that a careless generalization of precedence in the scientific cycle to precedence in time is absurd, as it would imply that temporally earlier findings cannot be used to contest the validity of later findings (see Kuperman, 2015 for a detailed analysis of this approach).

As far as we know, the term *megastudy*, referring to a collection of chronometric response data for many items, was coined by Seidenberg and Waters (1989), who had collected naming times for nearly 3000 words from 30 undergraduates at McGill University to examine how well their model captured item-level variance. Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995) used the McGill data in a study about the importance of rimes in reading and decided to make their own set of naming data on 1329 consonant–vowel–consonant (CVC) words available, thus allowing the testing of hypotheses on different sets of naming data. Spieler and Balota (1997) collected naming data from undergraduate students on 2870 words that were typically used to train computational models of reading. In a later study, Spieler and Balota (2000) collected naming data on the same set of items from adults aged over 70 and compared them with the data they had collected from the undergraduates. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) were the first to collect data using the lexical decision procedure, starting with 2902 words, again for younger and older adults. This was followed by the Balota et al. (2007) publication of the English Lexicon Project, in which both lexical decision and naming data were collected for 40,481 words from hundreds of participants (each responding to 3400 stimuli in lexical decision or 2000 items in naming) at different universities in the United States. Ferrand et al. (2010) published a similarly collected set of lexical decision reaction times for 38,840 French words. In the first lexical decision megastudy for Dutch, Keuleers, Diependaele, and Brysbaert (2010) tried a slightly different approach. Each of the 39 participants in their study responded to all 28,000 stimuli (14,000 words and 14,000 nonwords). A similar approach was taken with the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012). Data on 28,000 English words were obtained by testing 78 British participants at Royal Holloway University, each of them responding to one of two lists of 28,000 stimuli in lexical decision. A smaller project was initiated by Yap, Rickard Liow, Jalil, and Faizal (2010) who collected lexical decision and speeded naming for 1510 Malay words. Sze, Rickard Liow, and Yap (2014) moved megastudies in the domain of nonalphabetic writing systems, collecting lexical decision data for 2500 Chinese characters.

Recently, the megastudy approach has been extended to priming, where the focus is on covering different types of semantic and orthographical relatedness between prime and target rather than on covering a wide range of lexical characteristics. Hutchison et al. (2013) collected speeded naming and lexical decision data for 1661 target words, preceded by semantically related and unrelated primes at stimulus onset asynchronies (SOAs) of 200 ms and 1200 ms. Adelman et al. (2014) gathered masked priming lexical decision data on 420 target words, each of which was presented in 27 different priming conditions.

Although recognizing the importance of standard factorial designs, Balota, Yap, Hutchison, and Cortese (2012) identify a series of particularly important questions that megastudy data can address. First, with a large enough set of items, one can use a regression approach instead of attempting to select items for cells within a factorial design to test predictions from models. This avoids the potential implicit biases that may occur when one selects specific items for specific cells (see, for example, Forster, 2000). Baayen (2010) provides a nice discussion of misconceptions about the issue of causal inference via regression approaches in this context. Second, because computational models make predictions at the item level across a wide set of items, megastudy data can be informative regarding how well the models do overall, and compared to standard predictor variables (see, for example, Perry, Ziegler, & Zorzi, 2010; Spieler & Balota, 1997). Third, megastudy data not only allow one to quantify the amount of variance captured by a variable but also encourage one to examine the functional relationships between continuous predictors and dependent variables instead of demonstrating the presence or absence of a difference based on dichotomization of the predictors (see, for example, Adelman & Brown, 2008). Fourth, megastudy data are useful in identifying and testing novel variables (e.g., Levenshtein distance measures developed by Yarkoni, Balota, & Yap, 2008). Fifth, researchers are beginning to use megastudy data at the subject level to examine individual differences in the trade-offs of predictor variables (e.g., Yap, Balota, Sibley, & Ratcliff, 2012).

The present issue adds significantly to the accumulation of megastudy data and inferences that can be drawn from such datasets. Ernestus and Cutler (2015) report on an important project called BALDEY—the Biggest Auditory Lexical Decision Experiment Yet. They presented 20 native speakers of Dutch with 2780 spoken Dutch words and a matching number of pseudowords. In their paper, they describe the methodological considerations in setting up the study and investigate two questions that are central to a good initial understanding of the data. The first question is when, during the stimulus presentation, participants in their auditory lexical decision task decide whether a stimulus is a word. In contrast to visual presentation, the speech modality only gradually reveals the word, and so one must wait for the sequence of sounds to unravel to unambiguously know the intended word. At a certain point in the sequence only one word will be possible—this is called a word's uniqueness point. In the auditory lexical decision task, however, there is always the possibility that the sequence will continue into a pseudoword. Therefore, participants should wait until the very end for making their decision. Ernestus and Cutler show that this is indeed the case: Total stimulus duration is a better predictor of response time than different measures of the uniqueness point. Their second question is which word frequency measure best predicts response times in the auditory lexical decision task. They find that the SUBTLEX-NL subtitle-based word frequencies (Keuleers, Brysbaert, & New, 2010) better account for their auditory lexical decision data than do CELEX frequencies (Baayen et al., 1995) and frequencies based on a smaller corpus of spoken text (Oostdijk, 2002).

Cortese, Khanna, and Hacker (2010) first applied the megastudy approach to episodic recognition memory, using a set of 3000 monosyllabic words. Cortese, McCarty, and Schock (2015) follow their earlier study using 2897 disyllabic words as stimuli. Their findings show a similar pattern observed in the original dataset. Specifically, recognition memory performance is mostly dependent on words' imageability, length, orthographic similarity, and frequency. Cortese et al. (2015) examine some fundamental theoretical issues in this paper. Specifically, the data present a critical problem for theories of recognition memory performance that account for the mirror patterns in recognition memory. In general, variables that produce better recognition of old items also produce better correct rejection for new items. For example, low-frequency words produce higher hit rate and lower false-alarm rate than high-frequency words. However, Cortese et al. found in this large dataset that the number of times a word is correctly recognized (hits) is not correlated with the number of misidentifications when it is used as a distractor (false alarms). Such an observation has not been addressed at the item level in past studies, most likely due to the factorial approach most commonly used in episodic recognition memory research.

The study reported by Taikh, Hargreaves, Yap, and Pexman (2015) addresses the extent to which processing words and pictures involves the same semantic system. Their study is based on stimuli from the International Picture Naming Project (Székely et al., 2004). Taikh et al. use a semantic classification task on the pictures as well as on the words corresponding to the pictures, arguing that a picture naming task is not ideal to investigate semantic processing because it involves formal lexical characteristics, such as the sound, the frequency, and the lexical neighbourhood of the word. When they investigate how strongly these formal lexical characteristics and semantic characteristics (e.g., imageability, semantic diversity) predict the obtained decision latencies in the semantic picture classification task, they find no involvement of the formal lexical variables. In the same task with words, however, both formal lexical and semantic variables play a role. On the whole, the authors interpret their results in support of the hypothesis that there is privileged semantic access for pictures.

Chetail, Balota, Treiman, and Content (2015) report a study that extends some earlier work in French on the important role of consonant–vowel sequences on processing words. They investigated this pattern in English utilizing both data from the English Lexicon project and data from the British Lexicon project. In particular, they show that hiatus words, such as *client,* which has a vowel cluster spanning a syllable boundary, are named more slowly than words in which the number of vowel clusters matches the number of syllables, but that the effect diminishes with increasing word frequency. In lexical decision, low-frequency hiatus words are processed more slowly than control words, while high-frequency words are processed more quickly. The results highlight the importance of consonant–vowel structures on word processing, which fits well with earlier findings concerning the effect of letter transpositions. Chetail et al. offer the elegant explanation that initial recognition and reading aloud of a word may be aided by the expectation of a canonical consonant–vowel pattern. When words deviate from that pattern, however, processing suffers. The authors argue that their findings suggest that models of word processing should include an intermediate level of orthographic representations between letters and words. The extension of the original pattern observed in factorial studies in French to two megastudy databases clearly establishes this hiatus pattern.

Sze, Yap, and Rickard Liow (2015) set out to determine the contribution of a large number of orthographic, semantic, and phonological variables to the processing of Chinese characters, using the lexical decision data collected in the Chinese Lexicon Project (Sze, Rickard Liow, & Yap, 2014). The paper comprehensively summarizes the earlier research in this domain and then cleverly synthesizes it in an analysis that both fills gaps and identifies new problems. It is therefore a true milestone in research on Chinese visual word recognition. The study sheds light on the role of phonology in Chinese word recognition, with a regression analysis on megastudy data contradicting earlier findings in support of phonological processing (Ziegler, Tan, Perry, & Montant, 2000). A virtual experiment using similar items, however, led to a partial replication of Ziegler et al.'s (2000) findings. As Sze et al. (2015) argue, however, this should not be taken as support for the hypothesis, but rather as a weakness of the factorial approach, which (a) severely restricts the range of items—28 items in the factorial design versus 1560 in the regression; (b) is blind to all but the most simple distributional characteristics of the matching variables; and (c) fails to control for other variables that are included in the regression (see Kuperman, 2015, for further discussion of resampling issues from megastudy data). In addition, the study provides a benchmark of effects in the development of computational models of Chinese word recognition. Before this study, our knowledge of Chinese character processing came almost exclusively from a number of small-scale factorial experiments, each investigating one particular variable while using matching to control for the effect of some of the others. Matching greatly reduces the set of possible items, making it difficult to generalize. Importantly, Sze et al. (2015) also report an effect of number of meanings on lexical decision times. In earlier studies, no effect of this variable was found. Demonstrating that the effect disappears when number of meanings is treated as a factor with two levels, Sze et al. (2015) show that a reason for the earlier lack of an effect was probably due to unnecessary dichotomization. In addition, they observed an interesting case of variable suppression, with the effect of consistency only becoming visible in the presence of other variables.

THE USE OF TEXT CORPORA IN PSYCHOLINGUISTICS: FROM COUNTING WORDS TO EXTRACTING MEANING

The English word frequencies published by Thorndike and Lorge (1944) were first used by Howes and Solomon (1951) to demonstrate the fundamental effect that word frequency has on word recognition time. Later, the word frequencies published by Kučera and Francis (1967) became an important staple of any study using words (most likely the most common stimulus in experimental psychology) in English. Initially lagging behind, lexical databases with frequency measures specifically tailored to psycholinguistics became available for other languages, like French (Radeau, Mousty, & Content, 1990) and Dutch (Baayen et al., 1995). In recent years, the quality of word frequency measures has rapidly developed as a result of the ease with which corpora of film and television subtitles can be created. After New, Brysbaert, Veronis, and Pallier (2007) first published word frequencies based on French subtitles, Brysbaert and New (2009) demonstrated that English word frequencies based on film and television subtitles substantially outperformed word frequencies based on Kučera and Francis norms in predicting the behavioural data that had just been collected in the ELP (Balota et al., 2007). Following this, subtitle-based word frequencies quickly became available for Dutch (Keuleers, Brysbaert, & New, 2010), Chinese (Cai & Brysbaert, 2010), Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), German (Brysbaert et al., 2011), Spanish (Cuetos Vega, González Nosti, Barbón Gutiérrez, & Brysbaert, 2011), British English (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), Polish (Mandera, Keuleers, Wodniecka, & Brysbaert, 2014), and Portuguese (Soares et al., 2014).

If corpora can be used to derive count measures explaining a large part of the variance in word processing speed, one can ask whether these corpora can also be used to derive other lexical characteristics. In this issue, several papers deal with the use of distributional semantics models. Broadly speaking, the term refers to the idea that a word does not gain its meaning in isolation but that it can be derived from its context. In the definition of Marelli, Amenta, and Crepaldi (2015), the:

> approach is based on the assumption that the meaning of a word can be approximated by the way that word co-occurs with other words in the lexicon. … word meanings are represented as vectors that are derived from these co-occurrences. The more often two words tend to occur with the same set of other words (i.e., in similar contexts), the closer their vectors are and the more similar their meanings are considered to be.

Three papers in this issue independently explore the possibility that, starting with a small seed set of words with an affective or semantic rating, distributional semantics models can then be used to derive ratings for unseen words. The intuition behind the approach is that the more words are semantically similar, the higher the probability that they will have similar ratings in semantic or emotional norms. Semantic similarity could then be the basis for assigning ratings to words for which no norms have been collected. Executing the approach requires two elements with an advanced degree of technical complexity. The first element is a distributional semantics model itself, which represents the words occurring in a corpus as numerical vectors based on their co-occurrence with other words. The second element is a statistical model to extrapolate ratings based on those similarities. Bestgen and Vincze (2012) first proposed to build a semantic vector space based on latent semantic analysis (LSA; Landauer & Dumais, 1997) and to extrapolate the norms (semantic estimates) based on the average rating for the 30 closest neighbours in this semantic space. Using cross-validation, they reported relatively high correlations between original and extrapolated ratings for valence, arousal, dominance, concreteness, and imageability with this technique. With the same goal in mind, Recchia and Louwerse (2015) take a different approach to extrapolating affective norms for valence, arousal, and dominance, but with similar success. Westbury, Keith, Briesemeister, Hofmann, and Jacobs (2015) also examined valence and arousal to identify a critical set of basic emotions that have remarkable power in capturing the ratings in large datasets and lexical decision performance. A different approach is taken by Mandera, Keuleers, and Brysbaert (2015), who examined combinations of different semantic vector spaces with different extrapolation techniques on ratings, and their predictive power in capturing word recognition performance. Interestingly, this also paper points out some important limitations to consider when using these generated ratings.

Marelli et al. (2015) use a distributional semantics model in resolving a consistent but overlooked finding: In morphological priming lexical decision tasks, stems used in transparent conditions (*farm* in *farmer*) are recognized more quickly than stems used in opaque priming conditions (*fruit* in *fruitless*), regardless of whether the prime is related or unrelated. They attribute this to the consistency between the target's orthographic form and its meaning, or, in other words, how informative any part of these target words is about the meaning of the words in which it occurs. Using a distributional semantics model, they formulate an orthographic–semantic consistency measure. After showing that the difference between the two groups of stimuli from masked priming experiments also manifests itself in the unprimed lexical decision data from the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012), they show that there is a facilitatory effect of orthographic–semantic consistency and that it generalizes to a much larger set of words.

The output of a distributional semantics model can be highly dependent on the choice of model, its parameter settings, the materials that compose the corpus, and the way the corpus is processed. De Deyne, Verheyen, and Storms (2015) address how corpus size and annotation impact a semantic model's ability to fit human relatedness judgements. They also compare models built from text corpora to models built from large natural word association data. While their approach relies on graphs rather than vectors, this appears to be mostly a notational difference. In all cases, the models built from word associations fit the human judgement data better than the models built on text corpora. The authors' conclusion is that the size of the corpus is not a critical variable, a finding they take to be relevant for acquisition, as the large size of some corpora is not realistic for what children know. In their opinion, the more important aspect is the signal-to-noise ratio of the corpus, or, in other words, the quality of the raw material for extracting semantic relations.

The work that De Deyne, Verheyen, and Storms (2015) present highlights the difference between naturally occurring corpus data and data directly elicited from participants. At first sight, this may appear to be a difference between artificial knowledge and human knowledge. To understand why this is only a superficial difference, it is good to be reminded that corpora are in essence also behavioural data. They are the result of human linguistic production processes. Measures derived from corpora are a form of crowd-based measures, where the crowd consists of writers freely creating text on different topics. The important difference with what we usually call behavioural data is that they are the result of a natural output instead of being elicited by researchers.

In their paper in this issue, Keuleers, Stevens, Mandera, and Brysbaert (2015) discuss how data collected in an experiment with hundreds of thousands of participants can be used to address questions about the effects of age, multilingualism, and education on vocabulary size. Using the large amount of data on a very diverse population, they identify two important trends: First, the results show that the increase in vocabulary size throughout life and in multiple languages is very similar to the logarithmic growth of number of types with number of tokens observed in text corpora. Second, the vocabulary that multilinguals acquire via related languages seems to increase their first language (L1) vocabulary size and outweighs the loss caused by decreased exposure to L1. In their paper, the authors also address the relationship between the word frequency measure that can be derived from corpora and a new *word prevalence* measure—how many people know a word—that can be derived from crowd-based experiments. They demonstrate that these two measures complement each other in explaining word recognition speed, following two principles. "*Where the corpus is weak the crowd is strong*" refers to the fact that about half of the word types in a corpus will have a frequency of one. A word frequency measure can therefore not contribute to explaining variation in behavioural performance for those words. The number of people who know each of these words (word prevalence), however, can vary widely. "*Where the corpus is strong the crowd is weak*" means that words that are known by nearly everyone, hence having nearly no variation in word prevalence, can have a very large variation in word frequency. Keuleers et al. show that, when word frequency, length, and neighbourhood are also considered as predictors, word prevalence appears to be the best independent predictor of lexical decision reaction times in the Dutch Lexicon project. Crucially, in contrast to other word count measures that have been proposed (e.g., Adelman, Brown, & Quesada, 2006), word prevalence and word frequency have a relatively small correlation (.35), and they independently account for large parts of variance in the behavioural data.

METHODOLOGICAL CHALLENGES

Many of the papers in this issue present new and inventive methods of dealing with specific issues encountered during the investigations. Kuperman's (2015) paper addresses an important methodological problem directly: How do we properly replicate an experimental study with megastudy data? A straightforward approach is to select the same set of items as that in the original experiment and run item analyses on those items. However, this approach ignores that the set of items in the original experiment is a sample of items fitting particular selection criteria and that megastudy data may contain many more such samples. Kuperman proposes a bootstrapping method that takes advantage of this. By running many virtual replications of the original experiment, megastudy data allow for a more detailed picture. It gives us a distribution rather than point finding and also allows one to examine how a particular selection of items fits within the overall distribution. By specifying the selection criteria and drawing items according to the selection criteria many times, specific selection bias is avoided. Even in a regression approach, where the whole set of megastudy data could in principle be used at once, a bootstrapping approach where the analyses are run on many subsets of the data has the advantage that it protects against overfitting (see also Mandera, Keuleers, & Brysbaert, 2015, for the application of this principle in the context of random forest models).

Cortese, Hacker, Schock, and Santo (2015) address the question whether megastudy data—which are typically collected in long experimental sessions—can be trusted to give reliable results despite fatigue. They investigated the consistency of the effects of word length, word frequency, orthographic consistency, neighbourhood size, and imageability in a naming study with 2614 words by presenting 585 critical words at the beginning or at the end of the experiment. While they find a general effect of fatigue—participants' responses slow down during the course of the experiment—this did not affect the size or significance of the effects. The authors note that many of the issues that are raised concerning the reliability of megastudies (e.g., Sibley, Kello, & Seidenberg, 2009) are relieved by taking z-scores of reaction times (RTs) by participant and by block (see also Balota, Yap, Hutchison, & Cortese, 2012; Keuleers, Diependaele, & Brysbaert, 2010). New analysis techniques such as factor smooths in generalized additive models promise an even more powerful way to deal with these issues (Wood, 2006).

CONCLUSION

Our goal in generating a special issue was to provide readers with a window to the application of megastudies, crowdsourcing, and large datasets in psycholinguistics. We believe that the papers in this issue have successfully done this, and hopefully the readers will appreciate this considerable utility of these approaches. In our opinion, the wealth of newly available norms, response data, and methods to collect and apply these data provide psycholinguists considerably more power in developing models of language processing.

ORCID

Emmanuel Keuleers http://orcid.org/0000-0001-7304-7107

REFERENCES

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227. http://doi.org/10.1037/0033-295X.115.1.214

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. http://doi.org/10.1111/j.1467-9280.2006.01787.x

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., … Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*(4), 1052–1067. http://doi.org/10.3758/s13428-013-0442-y

Baayen, R. H. (2010). A real experiment is a factorial experiment. *The Mental Lexicon*, *5*(1), 149–157. doi:10.1075/ml.5.1.06baa

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM). Linguistic Data Consortium*. Philadelphia, PA: University of Pennsylvania.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. http://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*, 639–647. doi:10.3758/BF03200465

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods* (pp. 90–115). Hove, East Sussex: Psychology Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. http://doi.org/10.3758/s13428-012-0195-z

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424. http://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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*(4), 977–990. http://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 1–8.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6), e10729. http://doi.org/10.1371/journal.pone.0010729

Chetail, F., Balota, D., Treiman, R., & Content, A. (2015). What can megastudies tell us about the orthographic structure of English words? *The Quarterly Journal of Experimental Psychology*, 1–22. http://doi.org/10.1080/17470218.2014.963628

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. http://doi.org/10.1080/14640748108400805

Cortese, M. J., Hacker, S., Schock, J., & Santo, J. B. (2015). Is reading-aloud performance in megastudies systematically influenced by the list context? *The Quarterly Journal of Experimental Psychology*, 1–12. http://doi.org/10.1080/17470218.2014.974624

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, *18*(6), 595–609. doi:10.1080/09658211.2010.493892

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *The Quarterly Journal of Experimental Psychology*, 1–13. http://doi.org/10.1080/17470218.2014.945096

Cuetos Vega, F., González Nosti, M., Barbón Gutiérrez, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *DIALNET*. Retrieved from http://recopila.uniovi.es/dspace/handle/123456789/10272

De Deyne, S., Verheyen, S., & Storms, G. (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology*, 1–22. http://doi.org/10.1080/17470218.2014.994098

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Language Sciences*, *1*, 218. http://doi.org/10.3389/fpsyg.2010.00218

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F-X., … Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE*, *6*(9), e24974. http://doi.org/10.1371/journal.pone.0024974

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, 1–20. http://doi.org/10.1080/17470218.2014.984730

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*(2), 488–496. http://doi.org/10.3758/BRM.42.2.488

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*(7), 1109–1115. doi:10.3758/BF03211812

Haagen, C. H. (1949). Synonymity, vividness, familiarity, and association value ratings of 400 pairs of common adjectives. *Journal of Psychology*, *27*, 453–463. Retrieved from http://search.proquest.com/docview/1290585641/citation/F4D1C9100B804CD1PQ/3?accountid=11077

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, *41*(6), 401. doi:10.1037/h0056020

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*(4), 1099–1114. http://doi.org/10.3758/s13428-012-0304-z

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. http://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*. http://doi.org/10.3389/fpsyg.2010.00174

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*(1), 287–304. http://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, 1–28. http://doi.org/10.1080/17470218.2015.1022560

Kučera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *The Quarterly Journal of Experimental Psychology*, 1–18. http://doi.org/10.1080/17470218.2014.989865

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*. http://doi.org/10.3758/s13428-012-0210-4

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*. http://doi.org/10.1037/a0030859

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. doi:10.1037/0033-295X.104.2.211

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, 1–20. http://doi.org/10.1080/17470218.2014.988735

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2014). Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*. http://doi.org/10.3758/s13428-014-0489-4

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of orthography-semantics consistency on word recognition. *The Quarterly Journal of Experimental Psychology*, 1–13. http://doi.org/10.1080/17470218.2014.959709

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. http://doi.org/10.3758/s13428-011-0124-6

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., … Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 122–130). Retrieved from http://dl.acm.org/citation.cfm?id=1866715

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04), 661–677. http://doi.org/10.1017/S014271640707035X

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report*, *10*(10), 357–376.

Oostdijk, N. (2002). The design of the spoken Dutch Corpus. In P. Peters, P. Collins, & A. Smith (Eds.), *New frontiers of corpus research* (pp. 105–112). Amsterdam: Rodopi.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1p2), 1–25.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*(2), 106–151. http://doi.org/10.1016/j.cogpsych.2010.04.001

Radeau, M., Mousty, P., & Content, A. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'année Psychologique*, *90*(4), 551–566. http://doi.org/10.3406/psy.1990.29428

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, 1–15. http://doi.org/10.1080/17470218.2014.941296

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464. http://doi.org/10.2298/PSI1004441S

Seidenberg, M. S., & Waters, G. S. (1989). Reading words aloud-a mega study. Retrieved from http://philpapers.org/rec/SEIRWA

Sibley, D. E., Kello, C. T., & Seidenberg, M. S. (2009). Error, error everywhere: A look at megastudies of word reading. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1036–1041). Retrieved from http://cogmech.ucmerced.edu/pubs/SibleyETAL09-csproc.pdf

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., … Perea, M. (2014). On the advantages of word-frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *The Quarterly Journal of Experimental Psychology*, 1–41. http://doi.org/10.1080/17470218.2014.964271

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, 8, 411–416.

Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, *15*(2), 225–231. http://doi.org/10.1037//0882-7974.15.2.225

Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*(1), 263–273. http://doi.org/10.3758/s13428-013-0355-9

Sze, W. P., Yap, M. J., & Rickard Liow, S. J. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *The*

Quarterly Journal of Experimental Psychology, 1–30. http://doi.org/10.1080/17470218.2014.985234

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., … et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. doi:10.1016/j.jml.2004.03.002

Taikh, A., Hargreaves, I. S., Yap, M. J., & Pexman, P. M. (2015). Semantic classification of pictures and words. *The Quarterly Journal of Experimental Psychology*, 1–17. http://doi.org/10.1080/17470218.2014.975728

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College Press.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, *124*(2), 107–136. doi:10.1037/0096-3445.124.2.107

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. http://doi.org/10.1080/17470218.2013.850521

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. http://doi.org/10.3758/s13428-012-0314-x

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology*, 1–24. http://doi.org/10.1080/17470218.2014.970204

Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: CRC press.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. http://doi.org/10.1037/a0024177

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*(4), 992–1003. http://doi.org/10.3758/BRM.42.4.992

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. http://doi.org/10.3758/PBR.15.5.971

Ziegler, J. C., Tan, L. H., Perry, C., & Montant, M. (2000). Phonology matters: The phonological frequency effect in written Chinese. *Psychological Science*, *11*(3), 234–238. doi:10.1111/1467-9280.00247